Lowering the Crime Rate of Cities with Machine Learning

Matthew Lefkowitz

December 15, 2019

Introduction:

All city legislative bodies and police force aim to lower their city's crime rate. One suggestion to accomplish this goal is to regulate the types of businesses that populate each neighborhood in order to discover if a relationship between business type and crime could be found. The core concept is rooted in the theory that the types of venues in a neighborhood can influence local crime.

In order to explore this concept, this paper will focus on one city in particular, St. Louis, Missouri, as a way of investigating the effectiveness of this method. In 2018, the crime rate in St. Louis, Missouri was approximately 3 times higher than the U.S. Average, with 187 homicides in 2018 and 205 homicides in 2017 [1]. St. Louis was chosen in the analysis both for its high crime rate and open data policy.

In order to perform this analysis, the neighborhoods in St. Louis will be clustered using machine learning methods from data on crime and business type. The clusters will then be displayed on a map and analyzed to determine the key links between venue type and business type and discover whether or not we can predict the type of crime that will occur when there are specific venue types in a neighborhood.

Data:

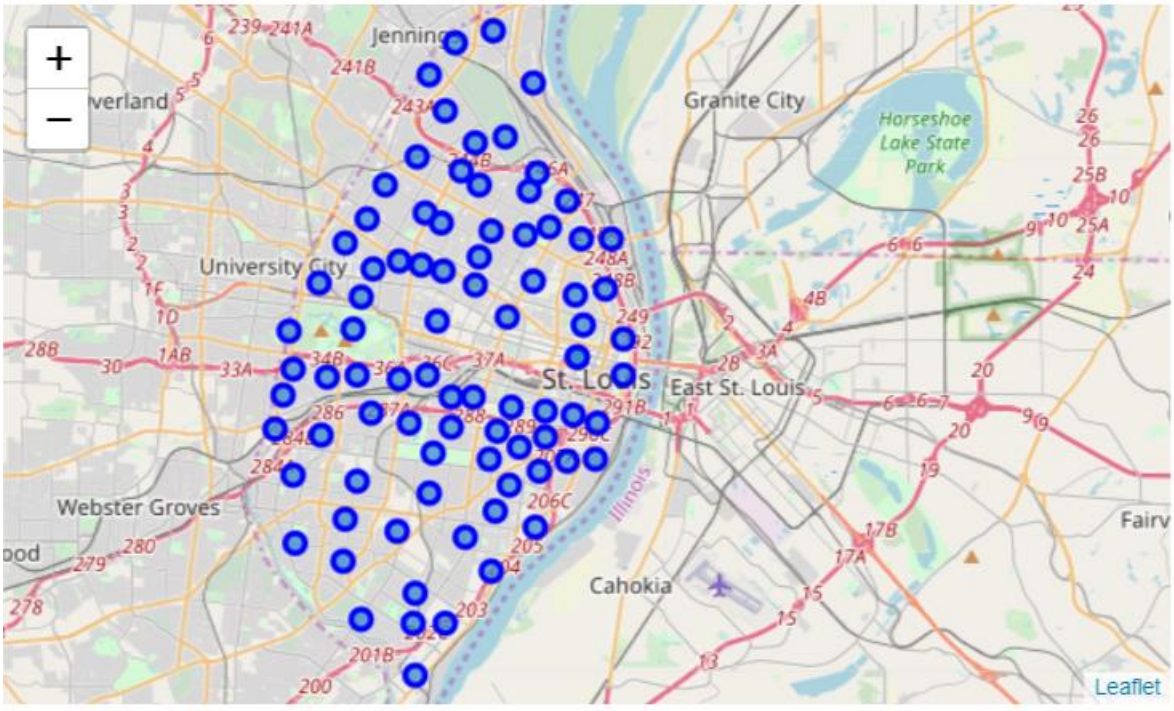
The data that was used to execute this project is publicly available crime data from St. Louis Missouri's website [2][3], along with data via Foursquare API [4] on the venues in the neighborhoods in St. Louis. Latitude and Longitude for each neighborhood was acquired via Google Maps [5]. The crime data contains all recorded crimes over the past year from November 2018 to November 2019. The dataset includes the type of crime, when the crime was recorded, and the neighborhood the crime took place in, among other variables. A full year of data was analyzed to account for outliers produced in seasonal crime differences based on a variety of factors such as weather and holidays.

Methodology:

First, the necessary packages were downloaded for the first code section, which were pandas and numpy. Next, the dataset for crime data was read into a dataframe from a csv file using pandas. A list of the neighborhoods and their corresponding number was also read into a dataframe neighborhood name could be added to the crime dataset via a merge. A groupby function was then used to get the count of the rows belonging to each neighborhood, which would show how many crimes were committed in each neighborhood. The number of crimes per neighborhood was then added back into the main dataframe with crime information and neighborhood names. The dataframe was created so it could be used for high level analysis on number of crimes per neighborhood. Another dataframe was created that only contained the crime description and the name of the neighborhood it occurred in. This was accomplished by dropping all other columns in the dataframe. This is the dataframe that was used for machine learning (ML) clustering.

Next, it was necessary to visualize the neighborhoods to make sure everything was pulling in correctly and in preparation for visualizing the clusters once the machine learning had been applied. Various visualization packages were imported including geopy for converting addresses to latitude and longitude, matplotlib for plotting, scikit-learn for kmeans clustering, and folium for map rendering.

A csv file with the latitude and longitude of St. Louis neighborhoods was read into a dataframe and merged that with the main dataframe with neighborhoods and crime numbers. Folium was then used to visualize the neighborhoods.



Once the neighborhoods had been visualized, data on business type was collected from Foursquare's API. The venues were read into a dataframe along with the neighborhoods they were associated with. The number of venues in each neighborhood was calculated with a groupby function. Onehot encoding and groupby functions were then used to get the frequency of each venue type occurring in each neighborhood. This dataframe of frequency was created for ML clustering.

	Neighborhood	ATM	Advertising Agency	American Restaurant	Antique Shop	Arcade	Art Gallery	Art Museum
0	Academy	0.0	0.0	0.00	0.0	0.000000	0.000000	0.0
1	Baden	0.0	0.0	0.25	0.0	0.000000	0.000000	0.0
2	Benton Park	0.0	0.0	0.04	0.0	0.000000	0.040000	0.0
3	Benton Park West	0.0	0.0	0.00	0.0	0.000000	0.055556	0.0
4	Bevo Mill	0.0	0.0	0.00	0.0	0.111111	0.000000	0.0
...	...	...	...	...	...	...	...	...
77	Walnut Park East	0.0	0.0	0.00	0.0	0.000000	0.000000	0.0
78	Walnut Park West	0.0	0.0	0.00	0.0	0.000000	0.000000	0.0
79	Wells-Goodfellow	0.0	0.0	0.00	0.0	0.000000	0.000000	0.0
80	West End	0.0	0.0	0.00	0.0	0.000000	0.000000	0.0
81	Wydown-Skinker	0.0	0.0	0.00	0.0	0.000000	0.000000	0.0

82 rows × 179 columns

The top venues by frequency were then found using sort functions and then placed in a dataframe with neighborhood and the most common venues in the columns adjacent. This was created so it could be used to evaluate ML clusters and to better understand the make-up of each neighborhood.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Academy	Outdoors & Recreation	Convenience Store	Video Store	Chinese Restaurant	Yoga Studio
1	Baden	American Restaurant	Pizza Place	Discount Store	Grocery Store	Yoga Studio
2	Benton Park	Fast Food Restaurant	Breakfast Spot	Sandwich Place	Bar	Dive Bar
3	Benton Park West	Mexican Restaurant	Pizza Place	Intersection	Convenience Store	Restaurant
4	Bevo Mill	Restaurant	Mexican Restaurant	Rugby Pitch	Lounge	Italian Restaurant

The same process was then used on the crime dataset to get the most common crime types per neighborhood.

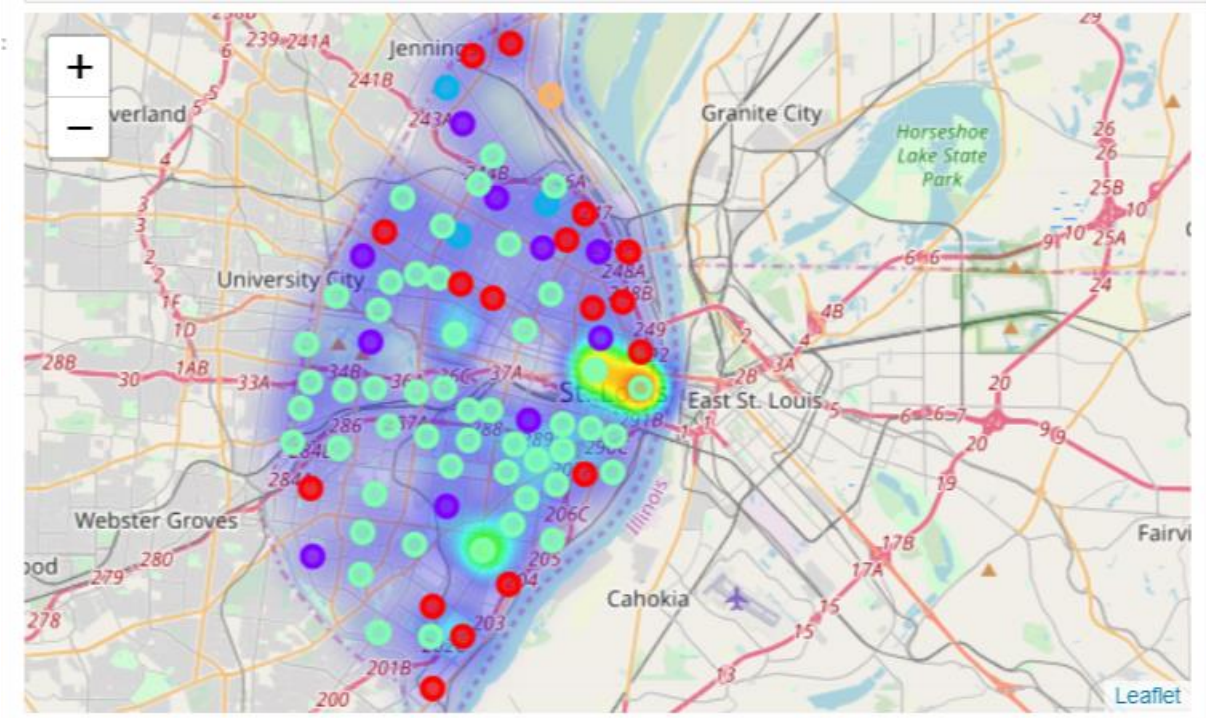
	Neighborhood	1st Most Common Crime	2nd Most Common Crime	3rd Most Common Crime	4th Most Common Crime
0	Academy	LEAVING SCENE OF ACCIDENT	DESTRUCTION OF PROPERTY-MALICIOUS/PRIV PROP	LARCENY-MTR VEH PARTS UNDER $500	STOLEN PROPERTY-BUYING,RECEIVING,POSSESSING,ET
1	Baden	LEAVING SCENE OF ACCIDENT	DESTRUCTION OF PROPERTY-MALICIOUS/PRIV PROP	AGG.ASSAULT-FIREARM/CITIZEN ADULT 1ST DEGREE	LARCENY-MTR VEH PARTS UNDER $500
2	Benton Park	LEAVING SCENE OF ACCIDENT	AUTO THEFT-PERM RETNT/UNRECOV OVER 48HR	DESTRUCTION OF PROPERTY-MALICIOUS/PRIV PROP	LARCENY-ALL OTHER UNDER $500
3	Benton Park West	LEAVING SCENE OF ACCIDENT	DESTRUCTION OF PROPERTY-MALICIOUS/PRIV PROP	AUTO THEFT-PERM RETNT/UNRECOV OVER 48HR	LARCENY-FROM MTR VEH UNDER $500

Once this was accomplished the two dataframes with venue type frequency and crime type frequency were then merged together for clustering. Kmeans clustering was used and the neighborhoods were grouped into 5 clusters. Each neighborhood was given a label 0-4 for which cluster they were a part of. The cluster label was added back into the dataframe with most common venue types per neighborhood, the number of crimes per neighborhood, and the longitude and latitude for each neighborhood. The neighborhoods were then visualized with folium, this time with the neighborhoods given a different color based on which cluster they were a part of. In addition, a heatmap function was added in using the number of crimes to determine the intensity of the heat visualized. Thus, the “hotter” areas are the areas with more crime, while the cooler areas are the areas with less crimes. This is to provide some additional context around the neighborhood clusters.

Each cluster was then examined based on the top crime types and venue types in the neighborhoods in that cluster. In addition, statistical data was calculated on crime data to get the average crimes committed, the min and max crimes per neighborhood, and the interquartile range.

Results:

The heatmap shows crime is mostly centered in downtown St. Louis, with sporadic incidents across the rest of the city. The clusters are also separated geographically, each neighborhood indicating the type of crime and business type is there based on the cluster it belongs to.



According to the descriptive statistics tables, all the clusters have a similar amount of average crimes committed over a year span aside from cluster 4, which reports about half as much as the rest. This is a good indication of proper clustering, as what separates the clusters is the venues and crime types, rather than the number of crimes committed.

The cluster with the lowest crime, cluster 4, has only one neighborhood and looks like it was in its own category because of the difference in venue make-up from the rest, with the highest reported venue being a waste management facility rather than a restaurant or similar venue. The crime is lower there, which could indicate that more industrial areas with perhaps less foot traffic is what is driving that number lower.

```
cluster_0.describe()
```

	Latitude	Longitude	Crime_y
count	17.000000	17.000000	17.000000
mean	38.636038	-90.234714	636.117647
std	0.051699	0.032016	409.915980
min	38.547032	-90.303774	282.000000
25%	38.600936	-90.252142	444.000000
50%	38.650537	-90.235219	481.000000
75%	38.667502	-90.209500	646.000000
max	38.719842	-90.190333	1869.000000

```
cluster_1.describe()
```

	Latitude	Longitude	Crime_y
count	10.000000	10.000000	10.000000
mean	38.644732	-90.247940	657.400000
std	0.036601	0.034269	421.391663
min	38.582563	-90.302508	45.000000
25%	38.624208	-90.276481	468.000000
50%	38.652039	-90.245538	547.500000
75%	38.664940	-90.224814	889.000000
max	38.698118	-90.204004	1494.000000

```
cluster_2.describe()
```

	Latitude	Longitude	Crime_y
count	3.000000	3.000000	3.000000
mean	38.684474	-90.244028	674.333333
std	0.021062	0.018932	139.378382
min	38.668268	-90.256935	577.000000
25%	38.672570	-90.254895	594.500000
50%	38.676872	-90.252855	612.000000
75%	38.692577	-90.237575	723.000000
max	38.708282	-90.222295	834.000000

```
cluster_3.describe()
```

	Latitude	Longitude	Crime_y
count	51.000000	51.000000	51.000000
mean	38.623935	-90.252067	617.235294
std	0.030099	0.031760	735.305939
min	38.561273	-90.309512	27.000000
25%	38.605810	-90.278385	198.000000
50%	38.621634	-90.254089	400.000000
75%	38.642671	-90.227812	649.000000
max	38.689750	-90.190283	3735.000000

```
cluster_4.describe()
```

	Latitude	Longitude	Crime_y
count	1.000000	1.000000	1.0
mean	38.705874	-90.220762	336.0
std	NaN	NaN	NaN
min	38.705874	-90.220762	336.0
25%	38.705874	-90.220762	336.0
50%	38.705874	-90.220762	336.0
75%	38.705874	-90.220762	336.0
max	38.705874	-90.220762	336.0

A feature table was created for each cluster:

Cluster	Most Common Venue Type(s)	Most Common Crime Type(s)
0	Bar/pub	Destruction of property, Leaving Scene of an Accident, Aggravated Assault
1	Park	Leaving Scene of and Accident, Destruction of Property
2	Gas Station, Convenience Store	Aggravated Assault, Weapons Violation
3	Restaurant, Hotel	Larceny
4	Waste Facility, Event Space	Larceny

Discussion:

Based on those results, one can successfully predict the types of crimes that were most likely to be committed by neighborhood with machine learning clustering. With these results the city of St. Louis can make legislative changes to lower the crime rate with this knowledge.

The legislative body can impose regulations on the amount of bars in a concentrated area to lower the rate of destruction of property and aggravated assault. They could provide infrastructure to assist in preventing aggravated assault and weapons violations at gas stations and convenience stores in ways such as bullet proof glass or automated pump use. They could make anti-theft notices in neighborhoods with many restaurants and in industrial areas and event spaces.

The police department can use this knowledge to devote different resources to different clusters of neighborhoods. In the cluster with more weapons related crimes they can send more units to patrol that area,

specifically around the related venues. In the cluster with more larceny around restaurants and hotels, they can place additional cameras and possibly employ recognition software to spot crimes in real time. Ultimately, knowing what types of crimes will happen can be an impactful resource for the St. Louis police department.

Conclusion:

The analysis in this report illustrates how one can use machine learning clustering to successfully group neighborhoods by crime and business type. These clusters can then be used to allow stakeholders to understand what crimes are likely to occur in different areas and take preventative actions to lower the crime rate. While the methodology was used on St. Louis, Missouri, it can be generalized to any other city with recorded venue and crime data. The analysis at this stage is not meant to be actionable, but rather is meant to demonstrate the general concept. In an actual use case, a deeper analysis should be performed to investigate each crime type and the neighborhood most strongly associated with it. Overall, what has been demonstrated is that cities worldwide can utilize machine learning for a more strategic deployment of resources that can greatly benefit their residents and create a safer future.

References:

1. "Crime in St. Louis", Wikipedia. https://en.wikipedia.org/wiki/Crime_in_St._Louis
2. "SMLMPD Downloadable Crime Files", Slmpd.org. http://www.slmpd.org/Crimereports.shtml
3. "Crime Data Frequently Asked Questions", St. Louis Metropolitan Police Department. http://www.slmpd.org/Crime/CrimeDataFrequentlyAskedQuestions.pdf
4. Foursquare API. https://developer.foursquare.com/
5. Google Maps. https://www.google.com/maps