Crime and Venues in Atlanta Neighborhoods

Rohan Chanani

November 29, 2020

1. **Introduction**
    a. **Background**
    Per Wikipedia.com, the Atlanta Metropolitan Area has a population of 6 million people and is the ninth largest metropolitan area in the United States by population. Atlanta also has a major issue with crime—98% of cities in the United States are safer than Atlanta. In fact, Atlanta's violent crime rate is double the national average. However, Atlanta's neighborhoods vary significantly in their amounts of crime, with some having much higher concentration than others. Most people want to avoid crime as much as possible, so information specific to neighborhoods about crime could be very useful.
    b. **Problem**
    I will first cluster the neighborhoods using their crime statistics. Then, I will use foursquare location data to build a classifier model that will predict the cluster of each neighborhood based off of its surrounding venues.
    c. **Target Audience**
    This model could be useful for both current and potential residents of Atlanta because it could help them understand different ways they can avoid crime. This could also be helpful to those living in other high crime cities for the same reason.
2. **Data Preparation**
    a. **Acquisition of Data**
        i. **Atlanta Police Department Data**
        I obtained the data about crimes that occured in Atlanta from the Atlanta Police Department at https://www.atlantapd.org/i-want-to/crime-data-downloads. Each row of the dataset is a crime that occured and the columns contain a large amount of information about the crime. Table 1 is an example of what some rows of the dataframe look like.

ii. **Foursquare Location Data**

I obtained data about crimes that occured in and around each neighborhood by making explore calls to the foursquare API with the latitude and longitude of each neighborhood and putting the results into a dataframe. Table 2 is an example of a dataframe created from the results of an API call

b. Data Cleaning

i. **Atlanta Police Department Data**

I needed four data points from each row in the Atlanta Police Department dataset: The neighborhood, the latitude, the longitude, and the type of crime. I split the dataset into two datasets—coords and counts. In the coords dataframe I had the average latitude and longitude of crimes that had occured in each neighborhood, and in the counts data frame I had the number of crimes that occured in each neighborhood, both the total number and the number of each crime type. I combined the two dataframes to get the full dataframe, with each row being a neighborhood and the columns being the latitude, longitude, total number of crimes, and the number of each type of crime(i.e. Larceny, burglary, homicide, etc.). Table 3 is the top of the full dataframe. Next, I made a dataframe that I would use for clustering. For the clustering dataframe, I dropped the latitude, longitude, and the name of each neighborhood, which left me with the total number of crimes for each neighborhood and the number of each type of crime for each neighborhood. I normalized the values for each type of crime by turning the values into a percent and then multiplying that percentage by the average of the total crimes column. Table 4 is the top of the clustering dataframe.

ii. **Foursquare Location Data**

The API calls to Foursquare returned json files from which I had to extract the relevant information. For each venue, I only really needed the category and the neighborhood the venue was in, but I wanted to get more information to help with the understanding of the dataset. I put the results into a dataframe with each row being a venue and the columns being the venue's neighborhood, name, latitude, longitude, category, and the neighborhood's longitude and latitude. I then did one-hot encoding to turn this dataframe into a dataframe in which each row was a neighborhood and each column

was the average value between 1 and 0 for each category within that neighborhood. The values for each category would be my feature set for later classification. Table 5 is the dataframe with these values.

Table 1:

ur te	Occur Time	Possible Date	Possible Time	Beat	Apartment Office Prefix	Apartment Number	Location	Shift Occurence	Location Type	UCR Literal	UCR #	IBR Code	Neighborhood	NPU	Latitude	Longi
)9-01	1145	2009-01-01	1148.0	411.0	NaN	NaN	2841 GREENBRIAR PKWY	Day Watch	8	LARCENY-NON VEHICLE	630	2303	Greenbriar	R	33.68845	-84.4
)9-01	1330	2009-01-01	1330.0	511.0	NaN	NaN	12 BROAD ST SW	Day Watch	9	LARCENY-NON VEHICLE	630	2303	Downtown	M	33.75320	-84.3
)9-01	1500	2009-01-01	1520.0	407.0	NaN	NaN	3500 MARTIN L KING JR DR SW	Unknown	8	LARCENY-NON VEHICLE	630	2303	Adamsville	H	33.75735	-84.5
)9-01	1450	2009-01-01	1510.0	210.0	NaN	NaN	3393 PEACHTREE RD NE	Evening Watch	8	LARCENY-NON VEHICLE	630	2303	Lenox	B	33.84676	-84.3
)9-01	1600	2009-01-01	1700.0	411.0	NaN	NaN	2841 GREENBRIAR PKWY SW	Unknown	8	LARCENY-NON VEHICLE	630	2303	Greenbriar	R	33.68677	-84.4

Table 2:

	name	categories	lat	lng
0	The Masquerade	Music Venue	33.751720	-84.389739
1	GSU Sports Arena	College Basketball Court	33.751735	-84.386328
2	Georgia Railroad Freight Depot	Event Space	33.751479	-84.388224
3	Willy's Mexicana Grill #22	Mexican Restaurant	33.751293	-84.385337
4	Jamrock Restaurant	Caribbean Restaurant	33.751554	-84.391356

Table 3:

Neighborhood	Count	LARCENY_NON_VEHICLE	LARCENY_FROM_VEHICLE	ROBBERY_PEDESTRIAN	ROBBERY_RESIDENCE	AUTO_THEFT	AGG_ASSAULT	BU
Adair Park	2012	399	440	170	14	343	198	
Adams Park	1504	310	344	92	15	202	114	
Adamsville	2798	603	699	206	22	419	304	
Almond Park	850	123	92	55	16	152	124	
Amal Heights	372	43	54	14	1	75	45	

Table 4:

	Count	LARCENY_NON_VEHICLE	LARCENY_FROM_VEHICLE	ROBBERY_PEDESTRIAN	ROBBERY_RESIDENCE	AUTO_THEFT	AGG_ASSAULT	BURGLARY_R
0	2012	269.759732	297.479403	114.935224	9.465254	231.898717	133.865732	
1	1504	280.379372	311.130658	83.209362	13.566744	182.698817	103.107253	
2	2798	293.158036	339.829962	100.150175	10.695650	203.703511	147.794433	
3	850	196.842280	147.231624	88.018906	25.605500	243.252249	198.442624	
4	372	157.238075	197.461768	51.193792	3.656699	274.252456	164.551474	

Table 5:

Neighborhood	ATM	Accessories Store	Adult Boutique	African Restaurant	Airport Terminal	American Restaurant	Animal Shelter	Antique Shop	Aquarium	...	Waste Facility	Weight Loss Center	Whisky Bar	Wine Bar	Wine Shop	Wings Joint	W
Adair Park	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
Adams Park	0.0	0.0	0.0	0.0	0.0	0.250000	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
Adamsville	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
Almond Park	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
Amal Heights	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
Wildwood (NPU-H)	0.0	0.0	0.0	0.0	0.0	0.250000	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.5	
Wisteria Gardens	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
Woodfield	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
Woodland Hills	0.0	0.0	0.0	0.0	0.0	0.060606	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
Wyngate	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	

3. Methodology

a. Exploratory Data Analysis

i. Choropleth Map

I decided to visualize the amount of crime in each neighborhood using a choropleth map to understand the density of crimes in different neighborhoods and to see if the geographical location should be considered a factor. Ultimately, outside of two outliers, Downtown and Midtown, the location didn't seem to be too large of a factor. The map is shown in Image 1.

ii. Scatter Plot Map

I made another map of Atlanta with each point being a neighborhood based off of the coords dataframe. I wanted to see the distribution of the neighborhoods and see if there was any pattern. Ultimately, there didn't seem to be any sort of pattern or clear factors associated with the distribution of neighborhoods. The scatter map is shown in Image 2.

b. Clustering

I decided to cluster the neighborhoods using KMeans to make the clustering as simple as possible because I thought this would make it easier when it came to classifying the neighborhoods. I decided to have four clusters, and they ended up being neighborhoods with Low, medium, high, and very high numbers of crime. The type of crime didn't have enough of a pattern to factor into the clustering. The very high cluster included Downtown and Midtown, the two outliers. Image 3 is a map that shows the different clusters.

c. **Classifying**

I decided to classify the models using K-Nearest-Neighbors, SVM, and a Decision tree and evaluate and tune them using cross validation with f1 weighted score, jaccard weighted score, and accuracy score.

4. **Results and Discussion**

None of the classifier models I used averaged above 0.6 on the evaluation metrics I used, which means they likely aren't accurate enough for use. There are a variety of reasons for why this might have happened. Many of the neighborhoods had little or no venues within 800m of their average, while some had over 100, which could create inconsistencies. There were over 300 different categories of venues, which may have made the model too specific and caused it to over fit. It's also possible that the types of venues in a neighborhood don't have an effect on that neighborhood's crime,

5. **Conclusion**

Although the classification models aren't accurate enough to predict the levels of crime in a neighborhood, the clusters are still helpful to help someone understand the safety of different Atlanta neighborhoods. To improve the models, I would try grouping the categories into larger, broader categories so there's a smaller number of categories. It would also be interesting to make a regression model that could predict the number of crimes in a neighborhood as opposed to the cluster that the neighborhood is in.

Image 1:

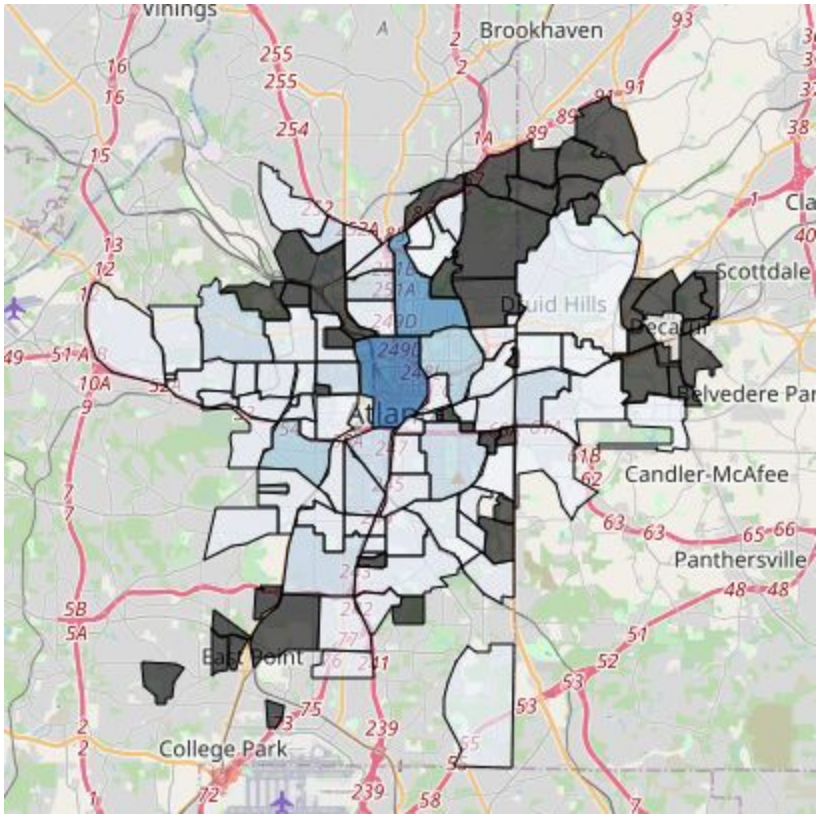


Image 2:

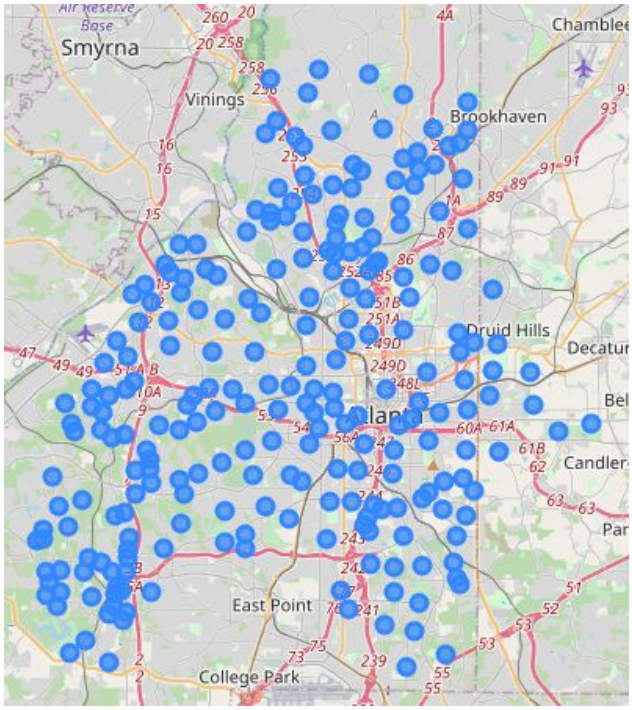
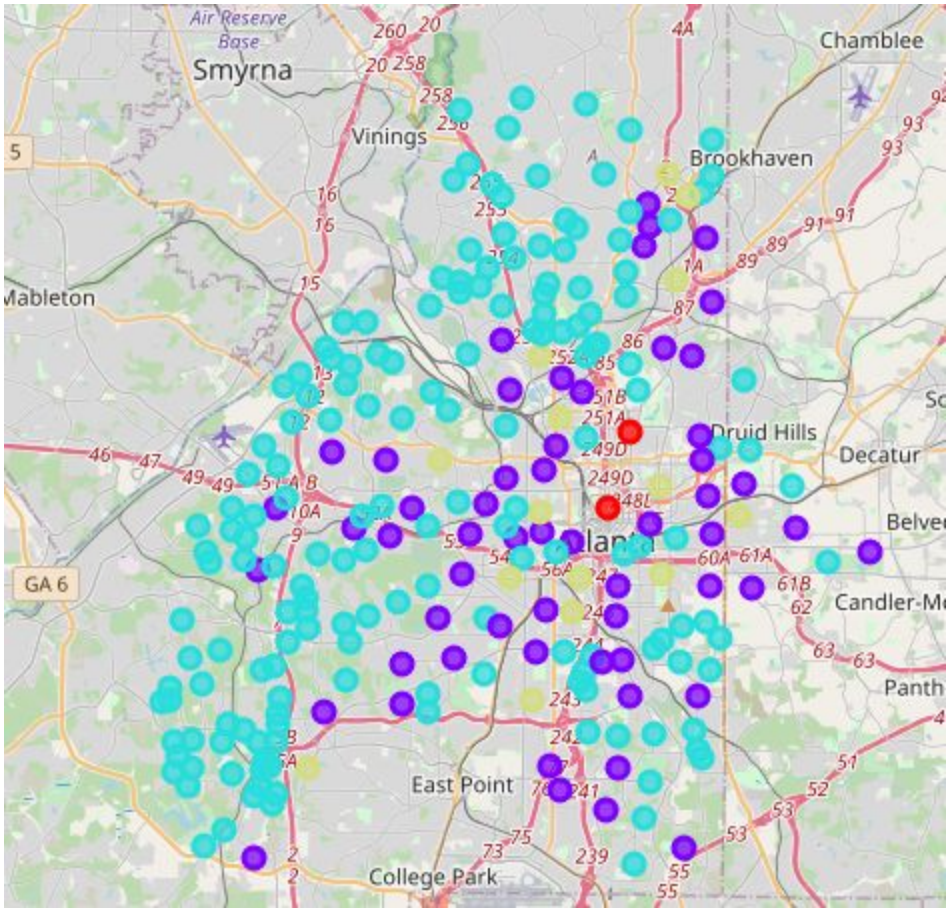


Image 3:


Air Reserve Base
Smyrna
Vinings
Chamblee
Brookhaven
Mableton
Druid Hills
Decatur
GA 6
East Point
College Park