2015 - 2022 Clustering of Crime Statistics in Nashville and Davidson County

Jacob Samar
jsamar1@vols.utk.edu
University of Tennessee
Knoxville, Tennessee, USA

ABSTRACT

The open sourcing of governmental data has opened up many different avenues for data analysis. It allows citizens and researchers the ability to identify trends or stories in the data that can inform individuals or policy makers of previously unknown information. One area of increased interest is around policing activity. There has been a general discontentment around police behavior in the United States over the last few years. As such I work to analysis the police actions within Davidson County over the last 7 years. This is enabled through a hierarchical clustering algorithm, HDBSCAN. This clustering allows for generating clusters at high density areas and ignoring low density areas, thus is able to generate identifiable crime hot-spots. This information can be used by citizens wishing to avoid more dangerous areas, and by policy makers to identify trends over the years for informing them on their legislative decisions.

KEYWORDS

crime, data analysis, clustering, hdbscan

ACM Reference Format:
Jacob Samar. 2021. 2015 - 2022 Clustering of Crime Statistics in Nashville and Davidson County. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

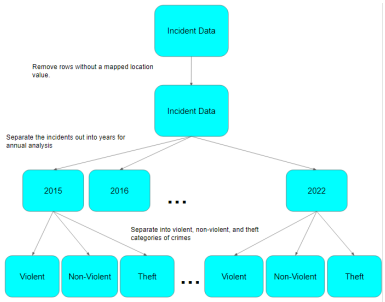
The number of incidents reported in a populated area such as Nashville is massive. Attempting to understand that data is a monumental task. Nothing can be gained by looking at the raw data. Simply plotting it onto a map of Nashville, even when reducing it to a single crime type, results in an entirely filled plot. There are no trends to identify because there is no separation of points. Initially, I worked to apply the k-means clustering algorithm to the dataset to draw out the patterns. However, this clustering fails to identify high-density areas vs low-density. It clusters each reported instance into a group, but those groups over no insight into were crime is happening more. Thus, a different clustering, HDBSCAN, was used[1]. HDBSCAN allows for ignoring the noise in a dataset, which makes it perfect for identifying crime hot-spots.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/1122445.1122456

2 METHODOLOGY

The dataset used is sourced from the Nashville government and includes every incident reported from 2015 to the present. An incident is defined as, "one or more crimes committed by an individual or a group of individuals acting in concert and at the same time and place" [1]. The dataset is first processed: removing rows with a null or zeroed location, separating each year of data, and finally dividing it into three subcategories (violent, non-violent, theft). The year 2022 obviously has incomplete data, which resulted in poor clustering, thus it is removed from the analysis completed. The



three categories are non-overlapping in the type of incidents they contain, except for aggravated robbery appearing in both violent and theft.

Once the data was processed, the clustering algorithm had to be prepared. Clustering is done using the hdbscan algorithm, which is a hierarchical variation on the DBScan algorithm that can form clusters of various densities, can account for noise in the dataset, and create arbitrarily shaped clusters. Since I am clustering latitude and longitude coordinates, I make use of the haversine metric, which is designed for working on latitude and longitude coordinates. hdbscan has 2 hyperparameters of interest:

(1) min cluster size: This is the minimum grouping size that will be considered a cluster
(2) min samples: A larger val of this leads to more points that are considered noise. Thus, higher values lead to more dense clusters

Min samples is important because the clustering of crime is intended to show only the more dangerous areas of Nashville. A higher min samples is selected because I wish to prevent clusters being formed on small pockets of crime and instead only identify areas were crime is very densely occurring.

HDBSCAN forms clusters of varying density and in arbitrary shapes, thus a Silhoutte score on the clustering is not the ideal way

[1] https://data.nashville.gov/Police/Metro-Nashville-Police-Department-Incidents/2u6v-ujjs

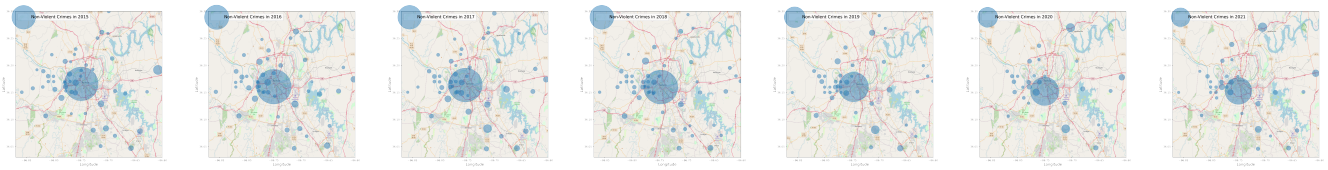


Figure 1: Non-Violent Incidents: 2015 - 2021

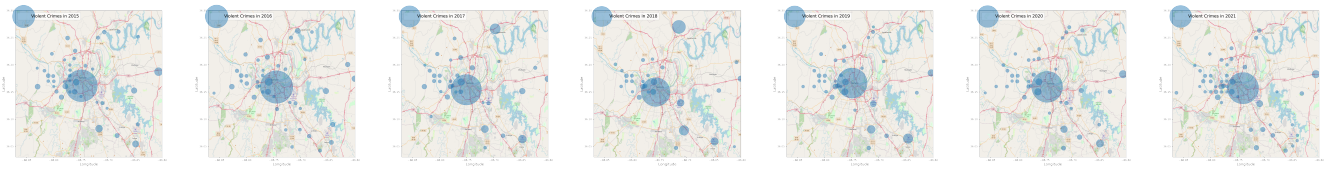


Figure 2: Violent Incidents: 2015 - 2021

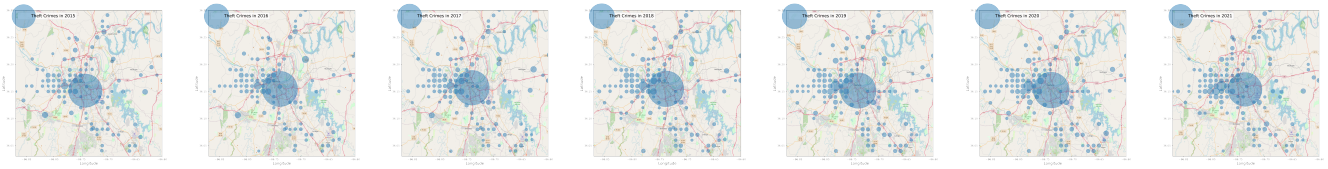


Figure 3: Theft Incidents: 2015 - 2021

to validate the clusters. Thus, I have instead performed a hyperparameter search over a range of potential values and verified the clusters visually. Lower values of min cluster size parameter results in a graphic covered in much smaller clusters and ultimately not providing much insight into the data. Lower values of min samples results in all points being clustered, also creating a massive amount of clusters and failing to show were higher densities of crime were occurring. 100, 100 were chosen as the parameters because higher values resulted in very large, all-encompassing clusters that offered no insight. 100, 100 was the nice middle ground of values that provides a reasonable number of clusters and cluster sizes. The hyperparameters that were searched are listed below:

- min cluster size: [10, 25, 50, 75, 100, 125, 150, 200]
- min samples: [4, 8, 10, 15, 20, 30, 50, 75, 100, 125, 150]

3 RESULTS

The largest cluster, which is present of downtown Nashville in each graphic, indicates crime mostly occurs in and around the south side of downtown. This is not surprising as it is the most frequented area in the Nashville area. The other clusters are where most of the insights can be drawn. For instance, analyzing all the graphics, you can see that all types of crimes are more prevalent around airports, TN State University, and Vanderbilt. Additionally, you can see more clusters, particularly of theft, to the west of Nashville, focused around an area known as Bell's Bend. Non-violent crimes have a trend of being located on or around marinas, campgrounds, and golf courses. Each of these areas have a plethora of unattended possessions, so intuitively this clustering makes sense. Interestingly, violent crimes are fairly sparse north of Nashville. This category is largely defined by clusters to the west and, less so, the south.

4 CONCLUSION

HDBSCAN is able to accurately identify areas of increased crime throughout the Nashville area (confirmed through matching large clusters to known areas of high crime)[2]. This clustering enables visual understanding of where particular types of crimes occur

[2]https://www.areavibes.com/nashville-tn/most-dangerous-neighborhoods/

in and around Nashville. Such information can be used by citizens, tourists, and policymakers alike. For visitors to Nashville, this offers insight on where to avoid (e.g. Hermitage Golf Course consistently has non-violent and theft crimes occurring despite little else around). By identifying the locations someone intends to go, they can understand the relative risk involved, whether with leaving possessions unattended or being aware of potentially dangerous individuals. Beyond that, this analysis can be used to inform and target legislative/police policy. Clearly the area west of Nashville and neighborhoods near airports are both more prone to all kinds of crime. Identifying this issue is the first step to understanding the trends of crime. Having identified problem areas, they can now delve deeper to understand why crime is more prevalent in these areas and hopefully target policy that will improve the area.

REFERENCES

[1] Ricardo JGB Campello, Davoud Moulavi, and J Sander. [n. d.]. Density-based clustering based on hierarchical density estimates.