PREDICTING CRIME HOTSPOTS AND ENHANCING PUBLIC SAFETY IN VIRGINIA

Alex Nguyen
hyv3jz

Jalen Wang
zac9nk

Colby Le
ncc9kn

January 23, 2024

1 Abstract

This project focuses on utilizing machine learning techniques to identify crime hotspots in Virginia, particularly in the Charlottesville area, in response to rising crime rates. The dataset was preprocessed to categorize crimes and relevant factors for clustering. K-means and DBSCAN clustering algorithms were applied, with K-means exploring various k values using inertia metrics and DBSCAN tuning epsilon and min sample values with silhouette scores. The project aims to enhance public safety by providing insights into crime patterns and hotspots, with ongoing efforts to fine-tune clustering techniques and incorporate temporal analysis. We hypothesize that if we are able to use spatial-temporal analysis on the Virginia crime data, we will be able to visualize clustered hot spots and forecast future crime rates in certain areas during certain times. This will allow for targeted crime prevention and efficient resource allocation that will hopefully improve Virginia's safety.

(Each group member contributed, with Colby handling data preprocessing and the Motivation and Methodology sections, Jalen constructing the K-means++ model and creating data visualizations, and Alex working on the DBSCAN model and the Preliminary Experiments and Next Steps sections.)

2 Motivation

According to the Charlottesville Police Department's 2023 annual report, the overall number of crimes this year by July compared to the same time last year saw an increase of 11.4% [5]. The prevalance of very violent crimes, such as sexual assault and homicide, have markedly jumped compared to last year, and as a whole crimes against persons and crimes against property also increased. However, certain crimes like drug offenses have decreased in prevalence. With the increased threat to public safety, the motivation behind this project is to cluster crimes and identify the patterns so that law enforcement can better understand the reasons behind why certain crimes are spiking or decreasing in prevalence. The overarching motivation behind this is to protect public safety in the Charlottesville area and to hopefully export these benefits to other localities in Virginia. Therefore, we do seek to frame this project as an application of machine learning rather than a theoretical experiment, because we want to identify real trends in the actual crime data and reap these benefits in public safety for the state of Virginia.

3 Data

The State Government of Virginia offers numerous datasets for practically any issue in the state. We are concerned about crime incident report data in Charlottesville, and there is a very comprehensive dataset titled "Charlottesville Crime Data" It can be found at `https://opendata.charlottesville.org/datasets/crime-data/explore`.

The dataset contains about 24.66 thousand incidents (rows) of data across 10 features. From a preliminary examination, the data is more than clean enough to use, as each row has detailed location data using a latitude and longitude feature as well as a street description, and at least a basic description of the crime. For example, there is a categorical feature that describes the type of crime (e.x. "simple assault" and "larceny"). We can expect to feature engineer several more columns, such as whether a crime is considered violent or not. This will give us more than enough to work with in the

clustering algorithm, and we can adjust the weights depending on the severity of each crime and also take location data into account for the clusters.

4 Related Work

Some related works include a study that focused on extracting patterns from spatial information. They used clustering and regression methods on generalized spatial data sets to identify categorized information [4]. Another research study that focused on crime prediction analysis looked at variables such as age, alcohol, media, etc. that could possibly affect crime rates. The factors listed did were seen to not have influenced much of the crime prediction, however, the conclusion in the study somewhat lacked a cohesive conclusion [2]. There have also been clustering algorithms that have attempted to glean patterns in crime so that authorities have more insight into how many crimes of a specific type may occur in an upcoming year. Specifically, they trained a k-means clustering algorithm on a crime dataset from England/Wales after cleaning the data and separating rows into the type of crime committed [1]. Therefore we know from the related works that there has been very successful research that has utilized k-means clustering algorithms on raw crime data, and research that has utilized clustering to generalize spatial trends. However, notably, there have not been any niche-related works that have combined the two applications using clustering, which is where we have room to create something new.

5 Methodology

The dataset and problem naturally fit the clustering problem. The algorithms applied in this problem include the k-means++ algorithm and the DBSCAN algorithm, both are implemented in the SKLearn machine-learning package. The reasons these two algorithms were selected are for their higher scalability than other clustering algorithms, given that the context of the problem may expand beyond Charlottesville, and the relative popularity of the models in conventional machine learning problems. It was also intended to have two clustering algorithms that compare examples using different metrics: k-means uses distance from the centroid, while DBSCAN uses distance to nearest point. The original presentation of the data also necessitated some data cleaning and feature engineering in the context of our problem. The data processing started by creating new categorical features: PersonalCrime, PropertyCrime, and NightCrime: these are boolean features that may designate certain examples of crimes that may involve personal violence (e.g. assault, homicide), destruction of property, or may have happened at certain times of the day. This way the model can weigh egregious crimes more heavily. Some extraneous features that would not be indicative of an overall pattern are dropped, such as the reporting officer and responsible agency. Scaling for numerical features and One-Hot-Encoding for categorical features are applied within a Pipeline. Various hyperparameter setups of the K-means and DBSCAN models are also tested: for K-means, this would be k (the number of clusters), while for DBSCAN this would be the epsilon value (maximum distance between two points to be considered a neighborhood), and minimum required samples to create a neighborhood. Our selected k-values reflected the number of officers who would be on duty at a given time, for example, if the Charlottesville Police Department could field 10 officers at a given time that is reflected in k = 10 clusters to police. Our most remarkable machine learning methodology in the first checkpoint is our data visualization, which plots the longitude and latitude of each crime on the Charlottesville map while displaying the clusters apportioned by the k-means++ and DBSCAN algorithms.

6 Experiments

The dataset provided three main types of independent variables for our experimental process; crime type, time of day, and geographical location based on longitude and latitude. During our data cleaning and pre-processing, we further categorized our types of crime into three main types including, personal, property, and night crime. Therefore, given our clustering techniques using K-means and DBSCAN, the dependent variables/insight that was achieved were possible crime clusters for hotspots.

Before discussing the results of our experiments (so far), it is also important to bring into light a major confounding factor that was not directly taken into account, but could play a role in the overall hotspots of where crimes were detected. This would include the current efforts of law enforcement and their current positions around the Charlottesville area. In order to control this variable, we processed the data so that the data was focused on crime patterns and did not include any information about the parole officers and their positions. The clustering results were driven by crime-related factors to capture crime incidents rather than police presence or activities.

For K-means clustering, k values of 3, 5, 9, and 15 were used to visualize clusters given the dataset. In order to compare the results, the metric of inertia was used to measure the compactness of the clusters. It was noticeable to say that at

k=3, the inertia value was the highest being 1157888. As k increased the value of inertia decreased with k=15 being the lowest of 63558. Given that this is was our initial values, it gives us good insight that k should not be a relatively low k-value so that the clusters become smaller and more tightly packed. However, it also brings insight that there is still the trade-off between fit and complexity where as k increases, the data will seem to appear closer to a respective centroid. Therefore, in order to compare the trade-off and also analyze the diminishing returns for the value of inertia, our next steps will take into account the elbow method in order to choose a future k value.

DBSCAN was also used for clustering with (epsilon, min sample) value pairs of (.05, 5), (.5, 5), and (.5, 10). In order to compare the DBSCAN trials, the metric of silhouette scores were used in order to evaluate how well-defined and separate the clusters are (higher score means clusters are more internally coherent). Given the DBSCAN trials with the value pairs previously stated, the silhouette scores for each trial were, -0.2239, -0.2983, and -.3352 respectively. A negative value represents that the clusters formed by the DBSCAN were not well-separated and there was a significant overlap between data points assigned to different clusters, meaning less distinct cluster boundaries. Although these results were not the most promising, DBSCAN could still give detailed insight on possible clusters if the parameters were extensively tuned.

K-means and DBSCAN clustering methods were both compared using silhouette scores where with K-means there was no distinct pattern, but an average silhouette score of 0.1438 between all trials. A slightly positive score shows that the K-means model was able to create clusters that were slightly better separated given the data points, however, this does not mean one model is better than the other. The decision was made to proceed with the k-means model because the inability to tune k aligns with the context of the problem. K-means models were constructed using different k values across two scenarios. Initially, the experiment fitted kmeans to the scaled data (using numerical and categorical features of location, time, and from k = 1 to k = 100. This yielded muddled clusters with no distinct clusters and models using k > 50 tended to be indistinguishable inertia wise, indicating a reduction in the marginal tightness of each subsequent group of clusters, so the experiment then proceeded to the first scenario with regard to this information. The first scenario entailed fitting kmeans to scaled location data (the longitude and latitude features) from k = 1 to k = 50 (these reduced k values reflect a realistic number of officers who could be on duty at a time). The second scenario performed compared the generated clusters during peak and lowest crime hours, that would be 2 PM and 5 AM respectively. Rows from the scaled training data on the longitude and latitude features which occured during these time periods were isolated, and again kmeans was fitted to this data from k = 1 to k = 30. For each scenario, across all k values, each model's inertia values (the sum of square distances of samples from their assigned centroids) were stored in a list. Those inertia values were then graphed against the k-values and utilized the elbow method to locate where the graphed inertia values begins to significantly slow, allowing the selection of a minimum suitable k (the minimum sensible number of officers who should be on duty at a time to serve well-fitted cluster boundaries). For the comparison scenario, the experiment then graphed the inertia of the 5 AM low crime set against the inertia of the 2 PM high crime set.

7 Results

Initially, we tested k-values of 5, 10, 20, 30, 40, 50 to assess the implications of different police force sizes on patrol efficiency across Charlottesville's 10.3 square miles. Our findings indicated that at k=5, the clusters were relatively large, suggesting that with only five officers on duty, each would need to cover a substantial area. This scenario, while providing a basic understanding of distribution, was not optimal for efficient patrolling.

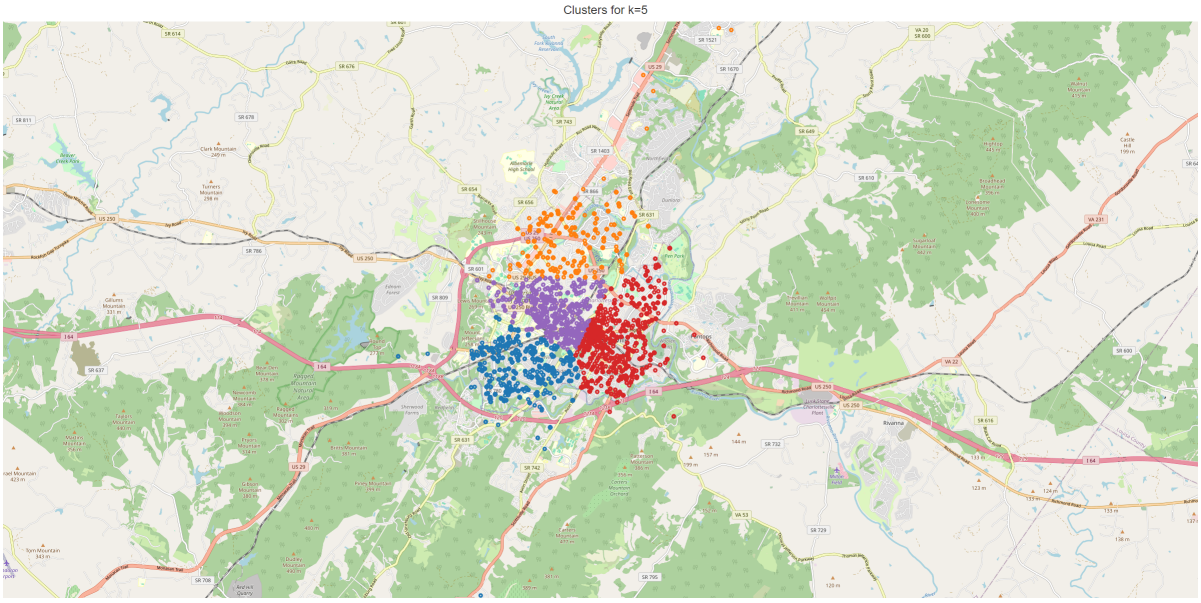

Figure 1: Clusters for k=5

To examine a more realistic scenario, we considered the situation with 15 officers on duty (k=15). The results showed significantly smaller clusters compared to k=5, indicating a more manageable area for each officer to patrol. This suggests that increasing the number of officers on duty enhances patrol efficiency by reducing the area each officer needs to cover.

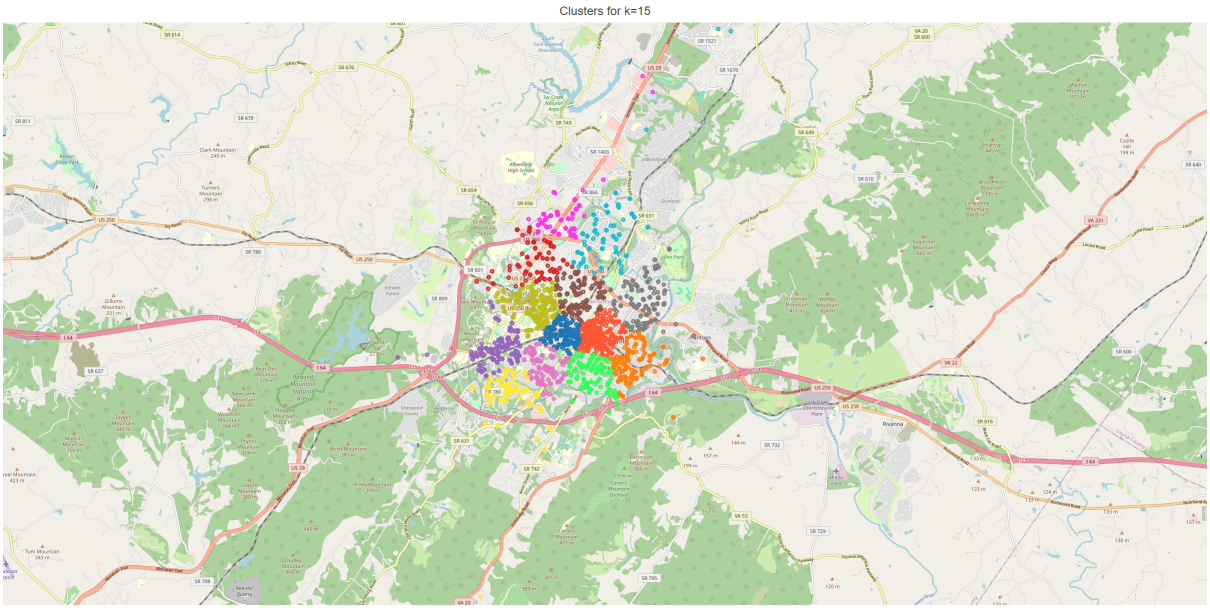

Figure 2: Clusters for k=15

The inertia values of the k-means clustering results played a crucial role in furthering our analysis. At k=5, inertia was 9794.12, reducing to 4484.32 at k=10, and further to 2708.78 at k=15. Using the elbow method, we observed a slowing

in inertia reduction between k=8 and k=15, around an inertia value of approximately 5000, indicating an optimal range for officer allocation in terms of cluster tightness and resource efficiency.

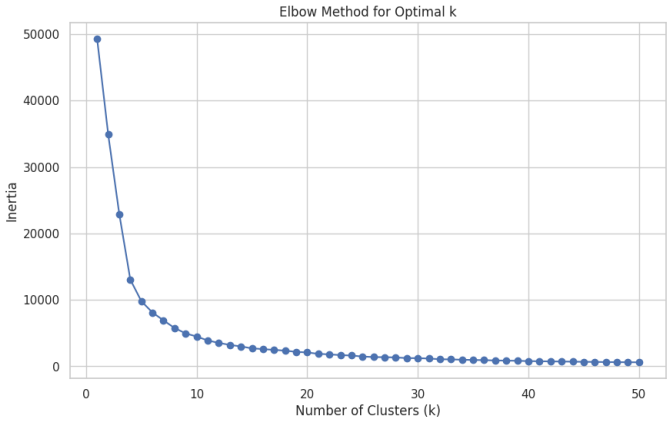

Figure 3: inertia values of scaled training rows. Elbow observed between k = 8 and k = 15, where inertia is roughly equal to 5000.

In the second scenario, we applied k-means clustering to subsets of the data representing crimes committed during two distinct times: 5 AM (low crime) and 2 PM (high crime). This analysis aimed to understand how crime patterns and optimal police force allocation might vary throughout the day. The models for the 2 PM data consistently showed higher inertia values than those for the 5 AM data, particularly at lower k-values. The 5 AM plot displayed an elbow at k=4, with inertia around 0.05, suggesting that four officers would be sufficient for effective coverage during these early morning hours. Conversely, the 2 PM plot indicated an elbow at k=10, with a similar inertia value, implying the need for a larger police presence during peak afternoon hours due to higher crime rates.

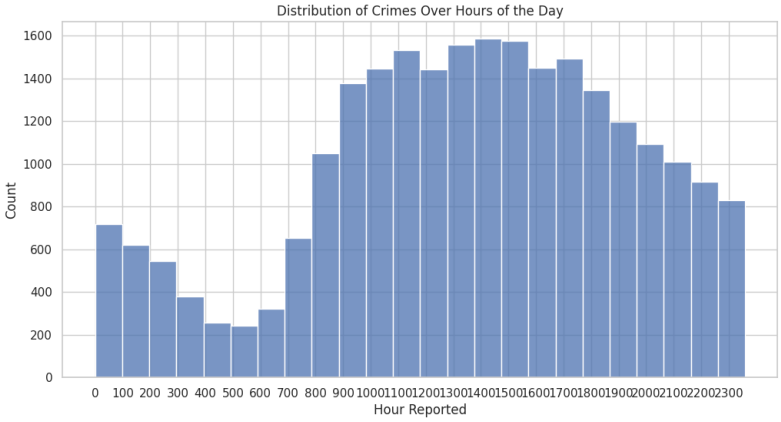

Figure 4: Histogram of crimes by hour they were reported. 5 AM is the slowest hour while 2 PM is the busiest.

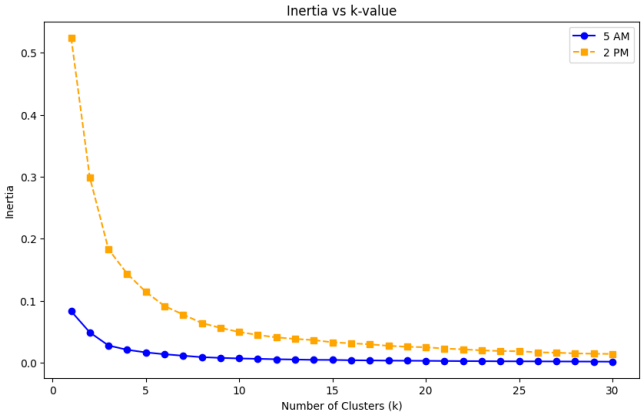

Figure 5: inertia values of 5 AM scaled training rows and 2 PM scaled training rows. Elbow observed around k = 10, where inertia was around 0.05.

8 Conclusion

What the results infer is that for the general case, the CPD should aim to field at least 8 or so officers in order to effectively police what the model has apportioned as well-fitted clusters or "districts." In reality this number can fluctuate, as the second scenario suggests that at slower hours such as 5 AM the CPD could get away with having as few as 4 officers on duty, while at 2 PM they could need 10 or more officers. Of course, it could only benefit them to have more officers on duty at all times, just at a lower marginal benefit in terms of necessity. The hypothesis absolutely is validated; machine learning can benefit public safety, as the k-means model was applied to real-world crime data in Charlottesville and yielded sensible numbers to advise the CPD on. Not only were the results important, the data visualization produced very useful and interesting graphs that re-draw the policing districts to something which is less arbitrary and more dynamic than current policing districts. The shortcoming in the experiment are mainly in the dataset. Though extensive, it only dates back several years and lacks sufficient detail to truly effectively sort training rows into clusters. Future planned steps will address these shortcomings, by attempting to stream live data from police reports and more instantaneous sources of information than the dataset, which is only periodically updated with new crime incidents. This experiment was created to improve public safety by advising police strategies, but it could be rehashed as a tool directly for the public. A mobile app could be created to allow users to view their own locations and avoid crime hotspots, and to directly submit live crime data.

9 Contributions

Each group member contributed, with Colby handling DBScan, data preprocessing, Motivation, Methodology, Results, and Conclusion sections, Jalen constructing the K-means++ model and creating data visualizations, and Alex working on the temporal analysis of the clusters, video, and writing the Experiments and Next Steps sections. All group members contributed voice lines to the video and helped with the final report.

References

[1] Agarwal, J., Nagpal, R., Sehgal, R. (2013). Crime analysis using k-means clustering. International Journal of Computer Applications, 83(4).

[2] L. Mookiah, W. Eberle, A. Siraj, Survey of crime analysis and prediction. In *The Twenty-Eighth International FLAIRS Conference*, 2015.

[3] Oliver, N. (2023, June 27). Violent crime is up in Richmond-area counties—And down in the city. Axios Richmond; Axios. https://www.axios.com/local/richmond/2023/06/27/richmond-violent-crime-rate-counties

[4] S. Shekhar, M.R. Evans, J.M. Kang, P. Mohan, Identifying patterns in spatial information: A survey of methods. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3), 193-214, 2011. `https://doi.org/10.1002/widm.25`

[5] Kochis, M. (2023). Charlottesville Police Department Annual Report. Charlottesville Police Department. https://www.charlottesville.gov/ArchiveCenter/ViewFile/Item/240