A Clustering Analysis of Policing in Chicago

Machine Learning for Public Policy, June 2021

Eric Chandler (echandler) and Jacob Lehr (jblehr)

Executive Summary

Background

This report examines the problem of discriminatory policing in Chicago. While there are many reports in the news and elsewhere describing the disproportionate impacts of policing on different communities, they are typically linear along racial or demographic lines, and relate to one outcome at a time. Our machine learning project focuses on clustering police activity in Chicago with the goal of identifying distinct groups that experience different types of policing. This will allow for both a more nuanced understanding of policing and for police departments, journalists, and the public to advocate for more specific reforms.

Data

We combine publicly available data from a variety of police, city, and federal sources. Specifically, we include data on police stops of pedestrians and vehicles, crimes, use of force incidents, and complaints filed by civilians. We also brought in demographic data from the U.S. Census. We merged the police-related and demographic data together at the police-beat-level (there are 277 in Chicago), and summarized the total count of each incident type by year for 2016-2019. These data allow us to merge previously siloed police activity data to get a comprehensive picture of policing in Chicago.

Machine Learning

To identify the best clustering model for this analysis, we first completed a principal component analysis of the relevant features, then searched across many combinations of different models and hyper parameters. Specifically, we tested the following models: K-Means, Gaussian Mixtures, DBSCAN, Hierarchical Agglomerative Clustering, and Spectral Clustering. Using a series of standard cluster-based evaluation metrics, we selected a version of the Hierarchical Agglomerative Clustering model as the model that identified the most promising clusters.

Evaluation and Results

Because this is an unsupervised learning project, we are not able to evaluate our models against a ground truth dataset. Instead, we compute several accuracy scores and validity checks to maximize the interpretability and usefulness of the best model. In addition, we also evaluated the chosen model on different time periods and demographic segments as a robustness check. We find high similarity between different years, and moderate similarity among different demographic-based models.

Overall, we determine clustering analysis, in this context, has limited practical benefits on this dataset over simpler linear descriptions.

Background and Overview of Solution

The Problem

A spate of recent police killings of Black Americans has put a spotlight on police brutality and racial bias in policing. The much-publicized killings of George Floyd, Breonna Taylor, Jacob Blake and others are not isolated incidents—police violence is a leading cause of death for young men, particularly young Black men.[1] Police stop data from across the country and at the federal, state, and local levels show that Black and Latino communities are consistently subject to a disproportionately high level of policing. In Illinois, the story is no different: Black and Latino drivers account for a disproportionately high share of both traffic stops and searches[2]; between 2005 and 2015 some 72 percent of recorded uses of force by the Chicago Police Department (the 2nd largest municipal police department in the country) were against Black residents.[3]

Clearly these are important problems, but they are not new problems; the CPD has been subject to a series of reform efforts since the 1970s. Most recently, in response to the police shooting of Laquan McDonald in October 2014, then mayor Rahm Emmanuel established a Police Accountability Task Force (PATF) and the U.S. Department of Justice opened a year-long investigation—both of which culminated in a number of reforms.[4] These reforms included establishing the Civilian Office of Police Accountability (COPA), a new "transparency policy" mandating the release of officer misconduct data and video, and the formation of a Community Policing Advisory Panel focused on improving community policing practices in Chicago.

In addition to the many news reports, we also identified recent work by the Urban Institute in Los Angeles to use clustering to better understand disproportionate impacts of policing.[5] The authors of this study further recommended similar analyses in other cities.

Our central question this report aims to answer is the following: There is known discrimination in policing. Can we identify additional useful distinctions in how groups face different types of policing?

Our Solution

Our goal is largely descriptive: by applying a clustering algorithm to a comprehensive dataset on Chicago police activity we hope to create a typography of police-civilian interactions. This type of typography would allow for a more nuanced understanding of both how police-community interaction differs across groups and the types of interactions that are more likely to lead to uses of force. Segmenting our data into time periods may also allow us to characterize how or whether policing in Chicago has changed over time in response to public pressure and reform.

We hope this analysis can be used to inform the debate around policing in Chicago and empower those in communities most effected by police violence. A greater understanding of how the police operate should allow for more targeted reforms.

---

[1] https://www.pnas.org/content/116/34/16793
[2] https://illinoistrafficstops.com/
[3] https://theintercept.com/2018/08/16/chicago-police-misconduct-racial-disparity/
[4] https://www.justice.gov/opa/file/925846/download
[5] https://www.urban.org/research/publication/catalyzing-policing-reform-data

Audience and Further Actions

The audience for the report (and model) includes several stakeholders. Because it is not a purely predictive model, it is not something that can be used to optimize police deployments. However, it can be used to better describe the impact of policing on communities, and better show the disproportionate impact in certain areas. These resulting clusters of different types of police activity should provide the following stakeholders with helpful information:

- ***Chicago Police Department***: Given that the department is actively seeking to both reduce actual and the appearance of discriminatory policing, providing simple groups of beats that experience a different type or level of policing can allow for the department to better target reforms to the community areas that need it the most.
- ***Journalists/news organizations***: The data presented and accessible in this report and model provide a new lens through which journalists can understand and report on policing. By clustering, we provide a grouping of Chicago neighborhoods which are sufficiently distinct, which can provide a more nuanced look at policing beyond simple linear demographic trends.

Data

Description

For this project, we combined several distinct datasets comprising several aspects of Chicago policing. The datasets are described in more detail in **Figure 1** below. We first aggregated, then merged the datasets at the police beat level, with summary data for each year between 2016 and 2019. This time period was the only time period during which we consistently had data across all datasets. We aggregated at the police beat level because we wanted our data to be spatially aggregated in a way that might reflect police decision-making, rather than rely on arbitrary Census or other city geographies. Police beat is the finest resolution data within the police geographical hierarchy.

In our final dataset, we have one row for each police beat and year, and the rest of the features are summarized as the count of the occurrence in that beat and year. For example, our measure of police stops is the total count of stops within that beat and year.

Because this is an unsupervised learning problem, we do not have labels for the data. Instead, our clustering process will assign labels as groups to better segment neighborhoods in Chicago that experience policing in a similar way.

Figure 1: Overview of data sources

Name	Source	Time period	Features Used
ISR (Investigatory Stop Report) Data	Chicago Police Department[6]	2016-2019	Total number of stops, total number of searches, total number of arrests, total number of stops by race (Black, White, Hispanic)
Use of Force Data	Chicago Police Department[7]	2016-2019	Total number of incidents, total number of incidents with/without a police weapon, total number of incidents by race (Black, White, Hispanic)

[6] https://home.chicagopolice.org/statistics-data/isr-data/
[7] https://home.chicagopolice.org/statistics-data/data-dashboards/use-of-force-dashboard/

Name	Source	Time period	Features Used
Complaints	City of Chicago[8]	2016-2019	Total number of complaints, total number of police shooting related complaints, total number of complaints by race (Black, White, Hispanic)
Crimes	City of Chicago[9]	2016-2019	Total number of crimes, total number of domestic crimes, total number of arrests, total number of crimes by race (Black, White, Hispanic)
Census Demographic Data	U.S. Census Bureau American Community Survey (ACS) [10]	2018 5-year estimates	Total population, median household income, percent of the total population by race (Black, White, Hispanic)

We followed several steps for each of the datasets to prepare them for analysis:

- ***Census Data:*** Census data was available at the block group level. Because the block group boundaries do not overlap exactly with our spatial unit of analysis of CPD Police Beat, we had to assign population to beats in some way. We created a weighted average population estimate for each beat by assuming that the portion of total area within a police beat corresponded to the total population of that census block group in the beat. For example, if 90 percent of a block group's area was within a beat, we assign 90 percent of the population to that beat.
- ***ISR Data***: The ISR data included many instances of record modification for a single stop, resulting in duplicate records. We selected only the data in the most recent record modification.
- ***Complaints:*** Many columns contained multi-valued categorical values that were often duplicated within and across rows, presumably indicating complaints on multiple counts, or against multiple officers, or filed by multiple people for the same incident.
- ***All datasets:*** For all data, we expanded categorical variables into dummy variables to summarize the relevant categories (e.g., generate counts by race). Out of hundreds of feature columns, we kept a flat dozen to simplify the interpretability of the resulting clusters.

Data Visualization

As part of our preparation for the modeling component of the project, we performed exploratory data analysis to better understand our data along temporal, spatial, and demographic lenses. A selection of these figures are provided in the Appendix.

Temporal Trends

As shown in **Figure A.1**, the datasets differ in their time trends. While total crimes and the incidence of use of force reports are relatively constant between 2016 and 2019, there is a significant upward trend in both total investigatory stops and complaints, beginning in 2017 and continuing through 2019.

[8] City of Chicago officer complaint data: https://data.cityofchicago.org/Public-Safety/COPA-Cases-By-Complainant-or-Subject/vnz2-rmie

[9] City of Chicago has reported crime data 2001-Apr 2021: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2.

[10] Census data accessed using their API: https://www.census.gov/data/developers.html

Spatial Trends

As shown in **Figure A.2**, there is clearly a spatial trend to the main features in our data. Specifically, especially at the police beat level of aggregation, we see similar beats grouped next to each other. This makes sense, as it is likely that beats are similar to nearby beats – the data just confirms this.

Demographic Trends

There are also demographic trends in the data. The plots in **Figure A.3** show one example of the clustered demographics we aim to further quantify and separate. The top plot shows what appears to be a weak and relatively un-clustered relationship between the total reported crime in a beat and the percent of all stops that involve a search. We might expect there to be a stronger relationship here if police are conducting more searches in high-crime areas. However, when we look at the percent of searches by the percent of the beat that is Black, we see a much more clustered dataset. This type of data in the two-dimensional space lends support that there may be meaningful clusters in the data at a higher-dimensional space.

Machine Learning and Details of Solution

Nature of Data

To simplify our analysis, we constructed our feature set to only include raw counts. All clustering methods rely on distance measures between pairs of observations, and therefore treat units as fungible between dimensions. This is not true of our data, where an incremental police shooting is qualitatively different than an incremental police arrest. Standard ways[11] to handle this include scaling up certain dimensions to have a higher cost or using multi-objective optimization to balance similarity on different attribute groups. Lacking a strong prior, we left our dimensions on equal weighting.

Raw counts also raised problems with features that represented rare events, like 'police shootings' and 'police use of force with weapon'. Since these incidents rarely occur, their distribution is only sparsely observed, and it looks noticeably quantized compared to the rest of the features. On one hand, we don't want our model to learn a spurious clustering on this distribution. On the other hand, these two variables happen to be qualitatively very meaningful, so generally we do hope they are important factors in the clustering.

Our data has meaningful spatial information, which is again qualitatively dissimilar from attribute distances. Again, lacking a strong prior weighting, we chose to drop explicit spatial information from our feature set.

Dimensionality Reduction

Seeing that most of our variables are strongly correlated, we ran a principal component analysis to reduce the data's dimensionality. We found that 80% of the total variance was explained by the first two components, and 95% by the first six (**Figure A.4**). Unfortunately, the data did not display any noticeable clustering in two-dimensional projections of eigenspace either (**Figure A.7**).

[11] https://geodacenter.github.io/workbook/9a_spatial1/lab9a.html

Models Considered

For the model selection step, we aggregated our data over all years and races to produce a complete representative training set. Our intent was to find a model that appropriately operates over the time period and demographic breakdowns conducted later.

Without better information on the shapes of our latent clusters, we ran several clustering models through a hyperparameter sweep: K-Means, Gaussian Mixtures, DBSCAN, Hierarchical Agglomerative Clustering, and Spectral Clustering. These methods have different strengths and failure cases; if a valid clustering exists, one of these methods should find it.

Figure 2: Overview of hyperparameter search

Model	Hyperparameters
K-Means	# of clusters
Gaussian Mix	# of clusters
Agglomerative	# of clusters, affinity metric, linkage
DBSCAN	epsilon, minimum # of neighbors, affinity metric
Spectral	affinity metric, # of clusters, minimum # of neighbors

- **# of clusters:** Prioritizing ease of explanation, we chose a half-dozen values between 2 and 10, and another half-dozen up to 80, covering the scale of Chicago's zip codes, wards, police districts, and community areas.
- **# of neighbors:** A natural scale for this parameter is the number of police beats, 272, divided by the number of clusters.
- **epsilon:** We used 6 log-spaced values spread between the minimum and maximum pairwise distances between observations.
- **# of principal components:** With each model, we clustered against four datasets: the top 2, 4, 6 principal components in eigenspace, all components in feature-space. In hindsight, it is hard to pick a 'best' number of components since distances in 2D aren't comparable to distances in 6D.

Model Selection

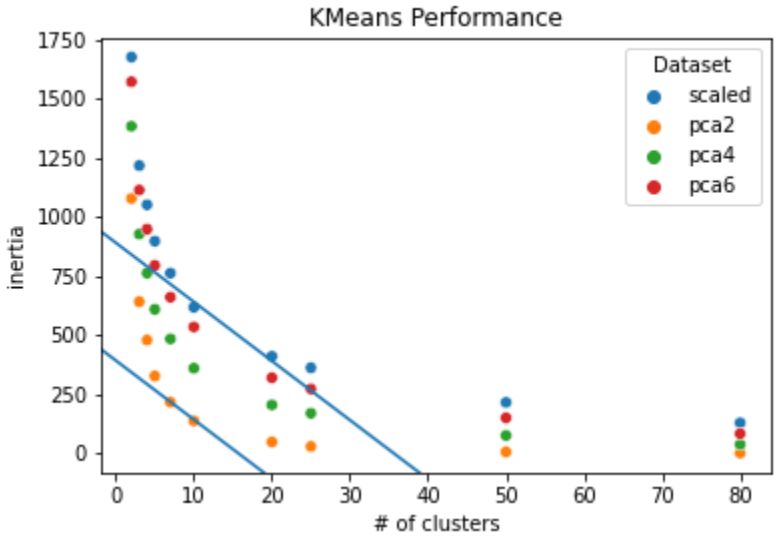
Model selection uses a two stage process, following the methodology in the Urban paper[12]. First, we chose the best hyperparameter set comparing results within models. Second, we chose the best among models.

Overall, no model clearly outcompeted everything else (**Figure A.6**). This likely isn't an artifact of the scope of our hyperparameter search, since within-model variance in performance isn't related to the number of hyperparameters tested. Instead, we attribute the across-models variance in performance to a lack of well-defined clusters in the data.

- **K-Means:** Our primary metric was inertia, which is the within-cluster sum of squared distance to centroid. This score approaches zero as the number of clusters increases, so the best point is identified when the marginal decrease in inertia becomes small (at 7-10 clusters in our case).

[12] https://www.urban.org/research/publication/catalyzing-policing-reform-data

Figure 3: Finding optimal K-Means



- **Gaussian Mixture:** We pick the hyperparameter set that minimizes the Bayesian Information Content score.
- **DBSCAN, Agglomerative, Spectral:** These methods tended to assign everything into the same cluster except for a few outliers. We considered models that assigned fewer than 80% of observations into the majority cluster and picked those that maximize the Silhouette score.

Model Comparison:

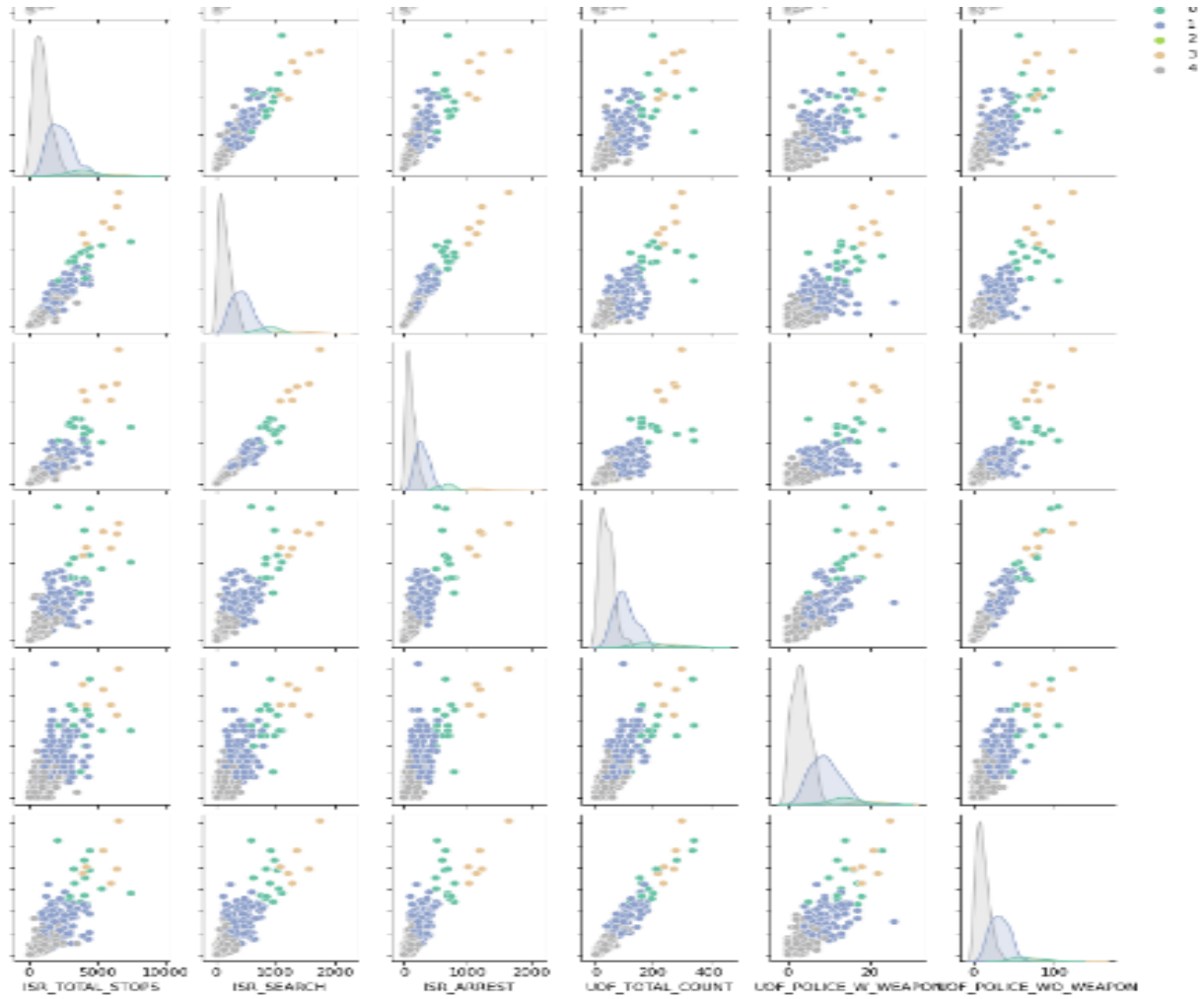
We compared models using the Davies-Bouldin, Calanski-Harabasz, and Silhouette scores – the three unsupervised clustering scores implemented in scikit-learn. Without a strong reason to prefer one metric over another, we ranked the models by their metric performance and chose the model with the lowest overall rank. We double-checked and found this heuristic to be robust accounting for the margin of performance.

Best Model:

We chose a hierarchical agglomerative clustering as our representative model. It breaks the features into two main groups and three groups of outliers. By contrast, the best K-Means model creates a low/med/high split of the main group and splits the outliers into four groups.

As shown in **Figure 4**, though this clustering is reasonable, the data doesn't exhibit obvious separation in any two-dimensional projection. With such high intra-model uncertainty, it is difficult to identify any particular model as the most truthful.

Figure 4: Clustering in Feature Space



Evaluation and Results

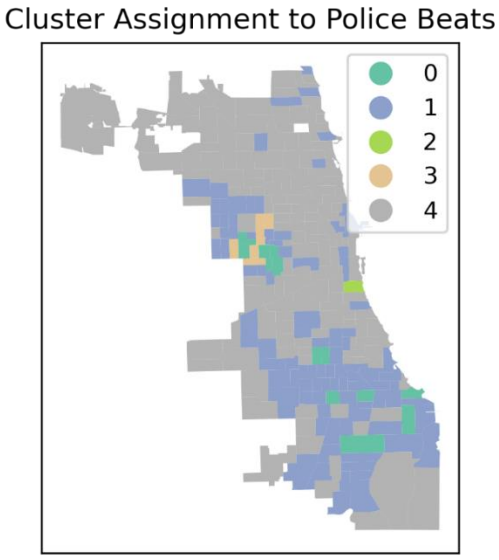
Feature Importance

We considered the feature importance of the resulting model. As shown in **Figure A.9**, we get varied values for feature importance based on the method used – logistic regression and decision trees. From the decision-tree-based method, we find that arrests (from the crime dataset) had the highest feature importance relative to the other features.

Summary of Produced Clusters

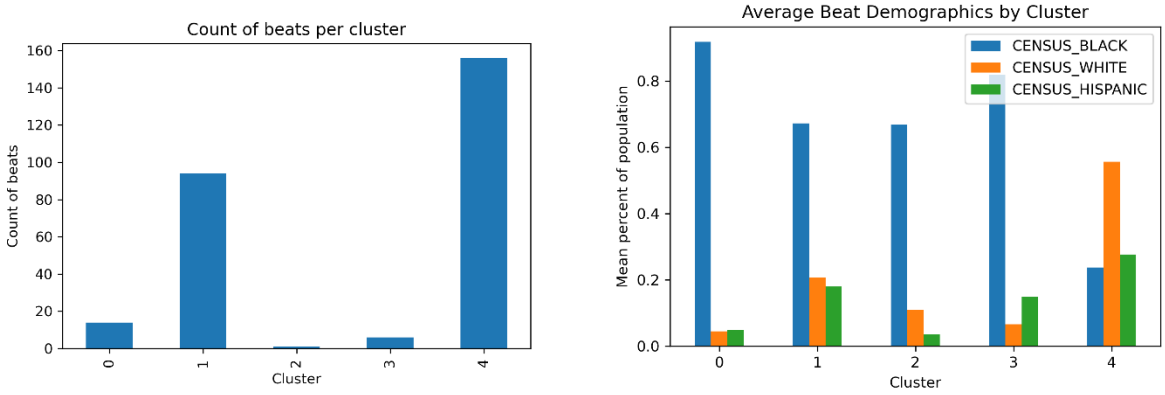
Since the nature of the clustering was descriptive and not predictive, one aspect of our evaluation involved examining the produced clusters.

Figure 5: Spatial Distribution of Clusters



As shown in **Figure 6** below, there are significant differences in the racial composition of different clusters. Since we did not explicitly use demographic features, it is striking that the majority-white beats were mostly clustered together and more nuance was assigned among the majority-black beats. A full summary of all relevant features is provided in **Figure A.8.**

Figure 6. Summary of Final Clustering by Police Beat



Comparison of Different Clusters

With a clustering methodology in hand, we extended our research question two ways. First, we ask whether policing has changed over time. Second, we ask whether our typology is robust across racial groups. To answer these questions, we developed the following algorithm to assess the similarity of pairs of clusterings:

1. Run our chosen hierarchical agglomerative model on a different subsets of the data (either a subset of features, or of rows/years)

2. For each clustering, construct an (nxn) adjacency matrix of police beats, where a pair of beats is adjacent if they are assigned to the same cluster. [13]
3. For each pair of adjacency matrices, calculate their cosine similarity, i.e. the number of common edges divided by the total number of edges.

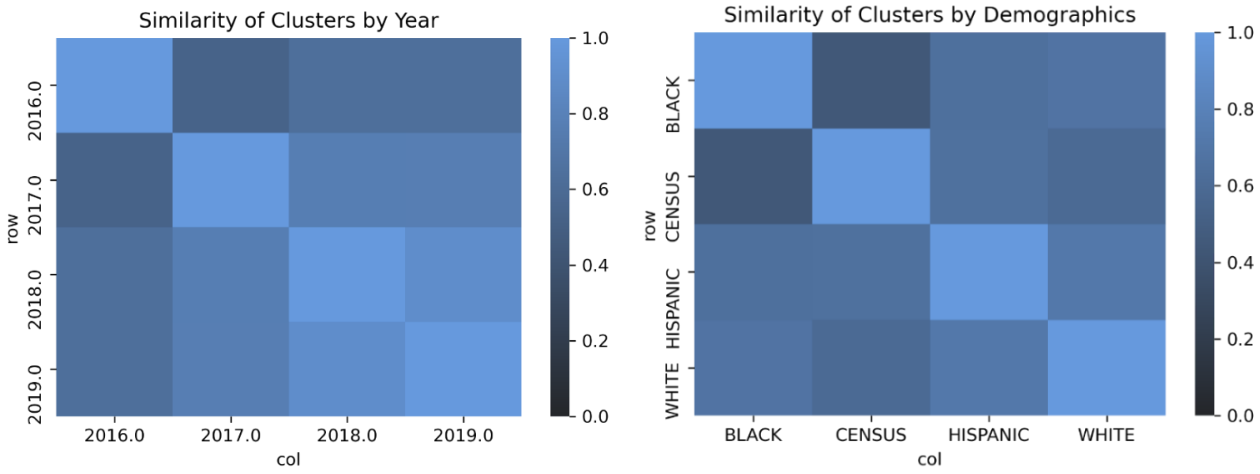
Comparison by year

When comparing across the years of analysis, we found that there is a moderate similarity in the clustering across years, showing that the clustering is not entirely random and does persist across years. The greatest dissimilarity was between years 2016 and 2017, possibly indicating a change in police practices at that time. **Figure 7** below shows the similarity between cluster models for each of the years between 2016 and 2019.

Comparison by demographics

To examine to what extent the clusters that we produced using our best clustering model are robust, we also compared its results to a clustering based on different holding out subsets of features. Specifically, we fit a new clustering model tuned to the same hyperparameters to only the features with demographic information attached to the counts of relevant activities (e.g., we include counts of stops of White, Black, or Hispanic individuals instead of raw counts). The maximum similarity between the two sets of clusters is 0.71. This is generally lower than the cross-year comparisons which was 0.89 at the highest. This might suggest that the clusters produced are driven more by the underlying demographics than by trends in policing[14].

Figure 7. Cosine Similarity of Clusters Produced over Different Cluster Models



[13] Where n is the number of police beats.

[14] Although, these demographic clusterings were based on a 3D feature-space instead of the 6D yearly featurespace and therefore encode less information.

Appendix

Figure A.1 Trends in Features Between 2016 and 2019

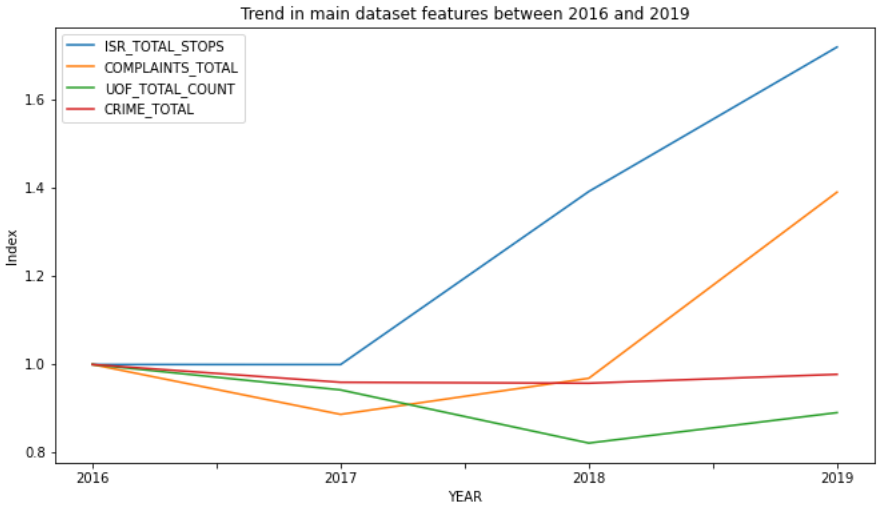


Figure A.2 Spatial Incidence of Main Features

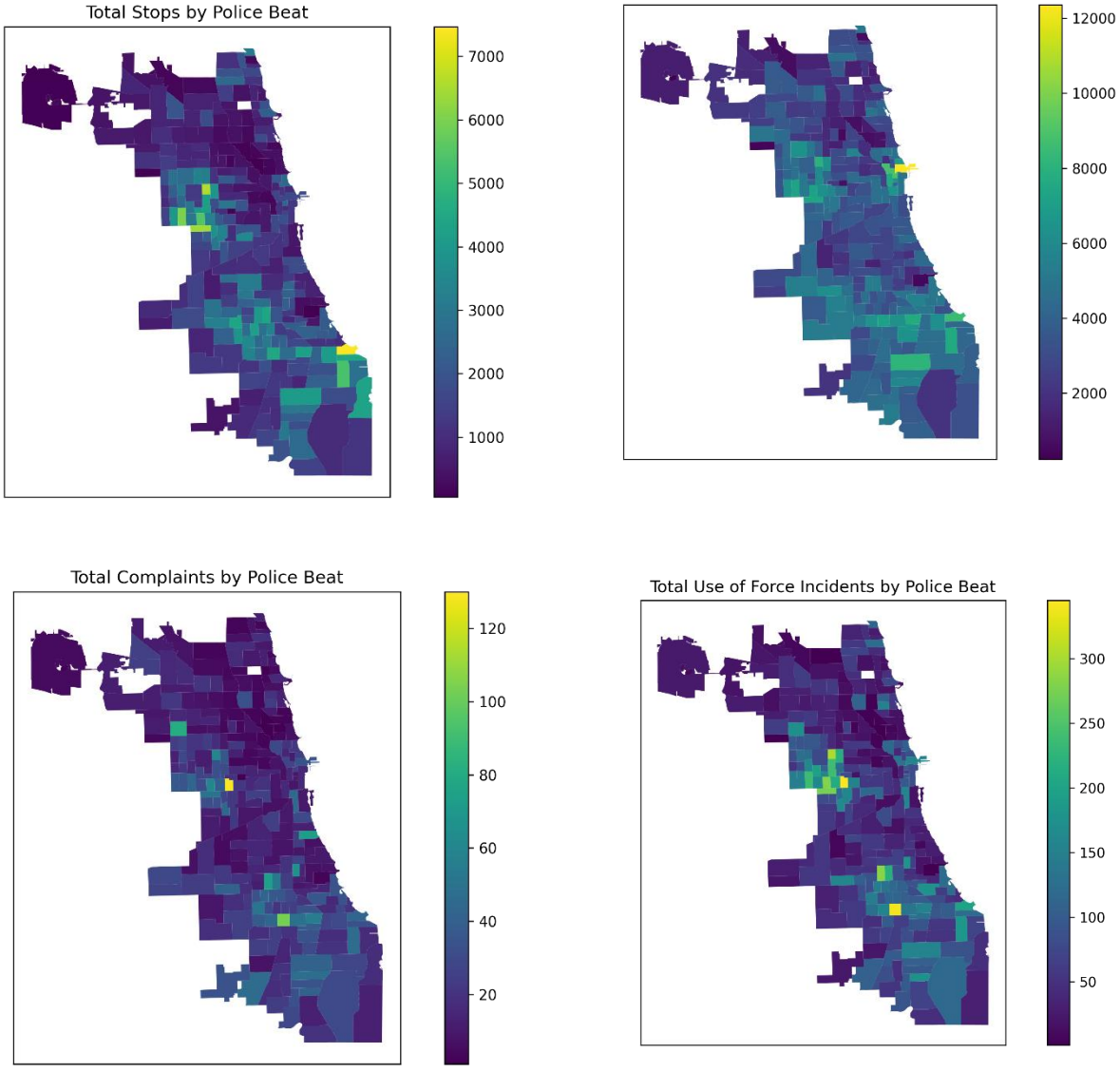


Figure A.3.1. Percent of Stops with Searches

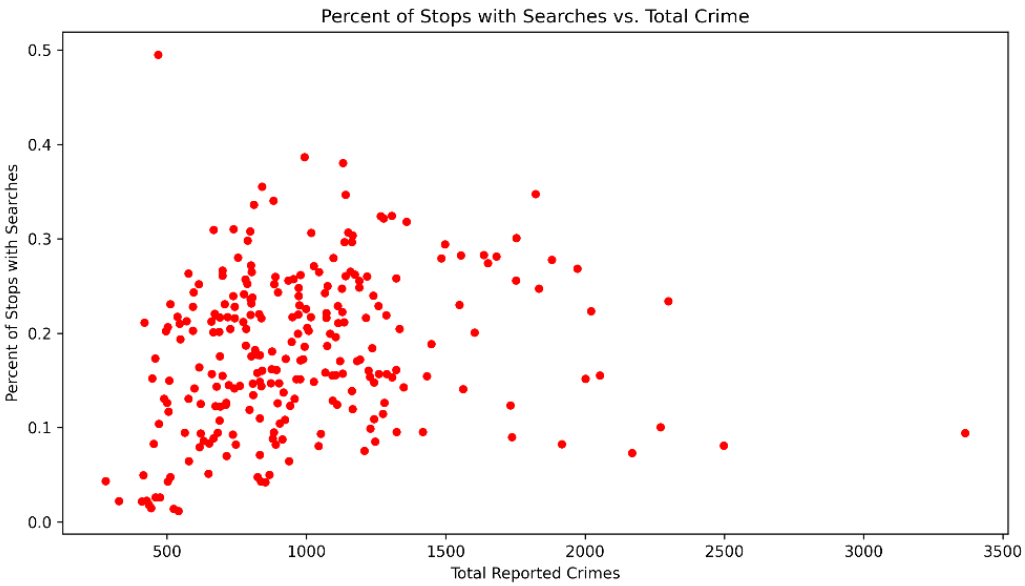


Figure A.3.2. Percent of Stops with Searches

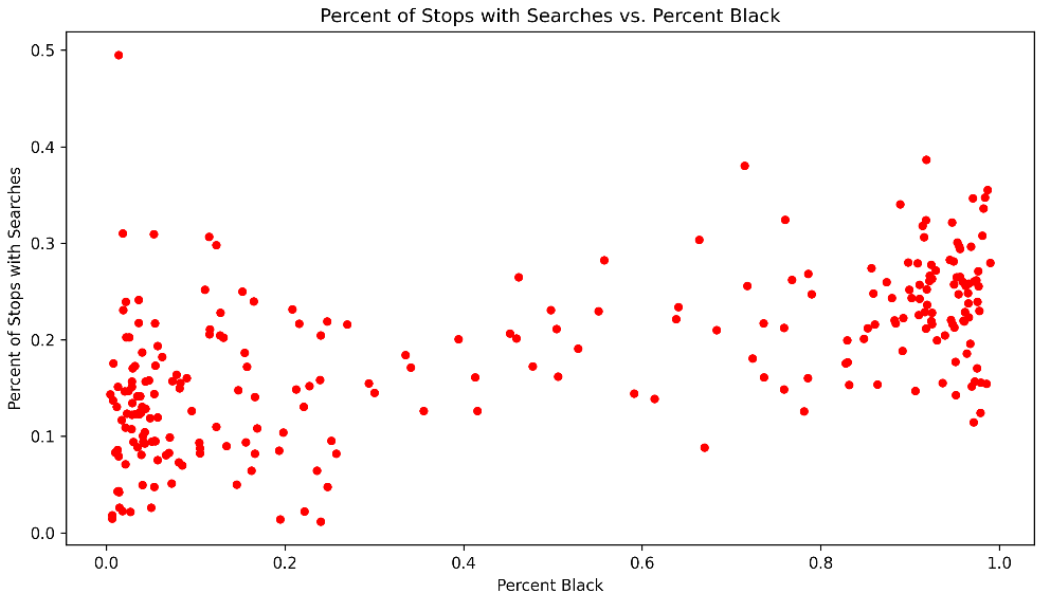


Figure A.4. Principal Component Percent of Explained Variance

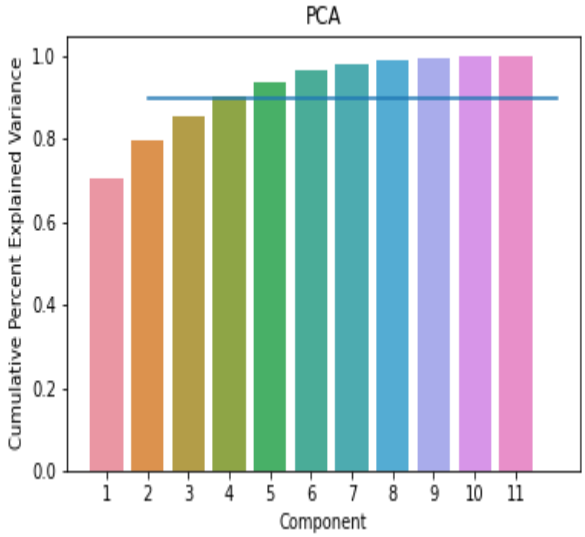


Figure A.5. Principal Components in Feature Space

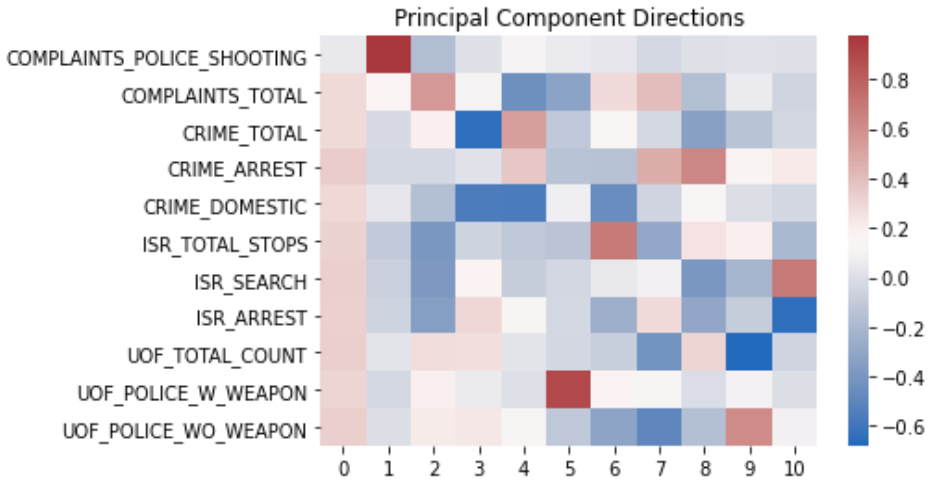


Figure A.5. Full Hyperparameter Search Specification

Model	# clusters	affinity	linkage	epsilon	# neighbors	# principal components
K-Means	[2,3,4,5,7,10, 20,25,50,80]	n/a	n/a	n/a	n/a	[2,4,6,11]
Agglomerative	[2,3,4,5,7,10, 20,25,50,80]	[euclidean, cosine]	[complete, average, single]	n/a	n/a	[2,4,6,11]
DBSCAN	n/a	[euclidean, cosine]	n/a	[0.056, 0.13, 0.31, 0.73, 1.7, 4.0]	[136, 91, 68, 55, 39, 28, 14, 11, 6, 4]	[2,4,6,11]
Gaussian Mix	[2,3,4,5,7,10, 20,25,50,80]	n/a	n/a	n/a	n/a	[2,4,6,11]
Spectral	[2,3,4,5,7,10, 20,25,50,80]	[radial basis, nearest neighbors]	n/a	n/a	[136, 91, 68, 55, 39, 28, 14, 11, 6, 4]	[2,4,6,11]

Figure A.6. Overall Model Performance Variance

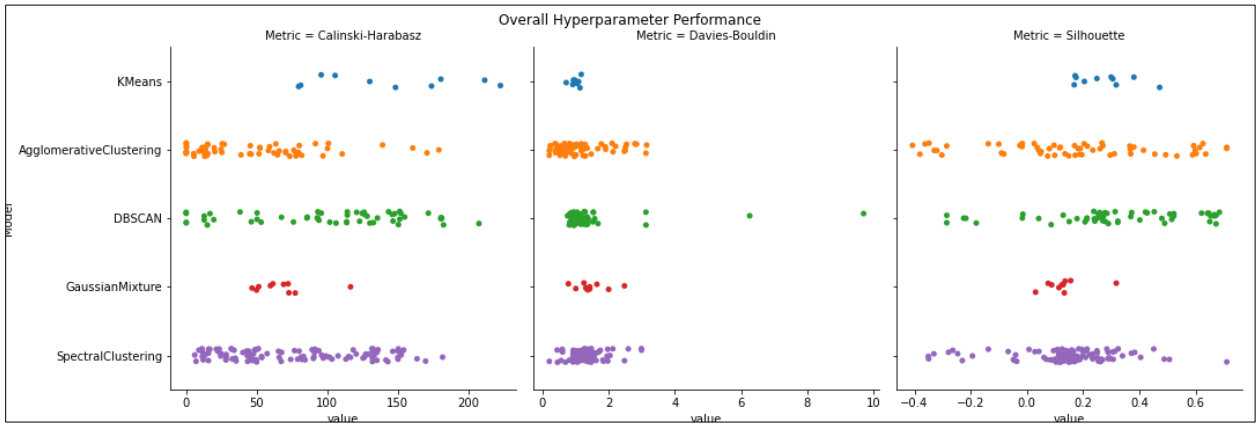


Figure A.7. Clusters in Reduced Eigenspace

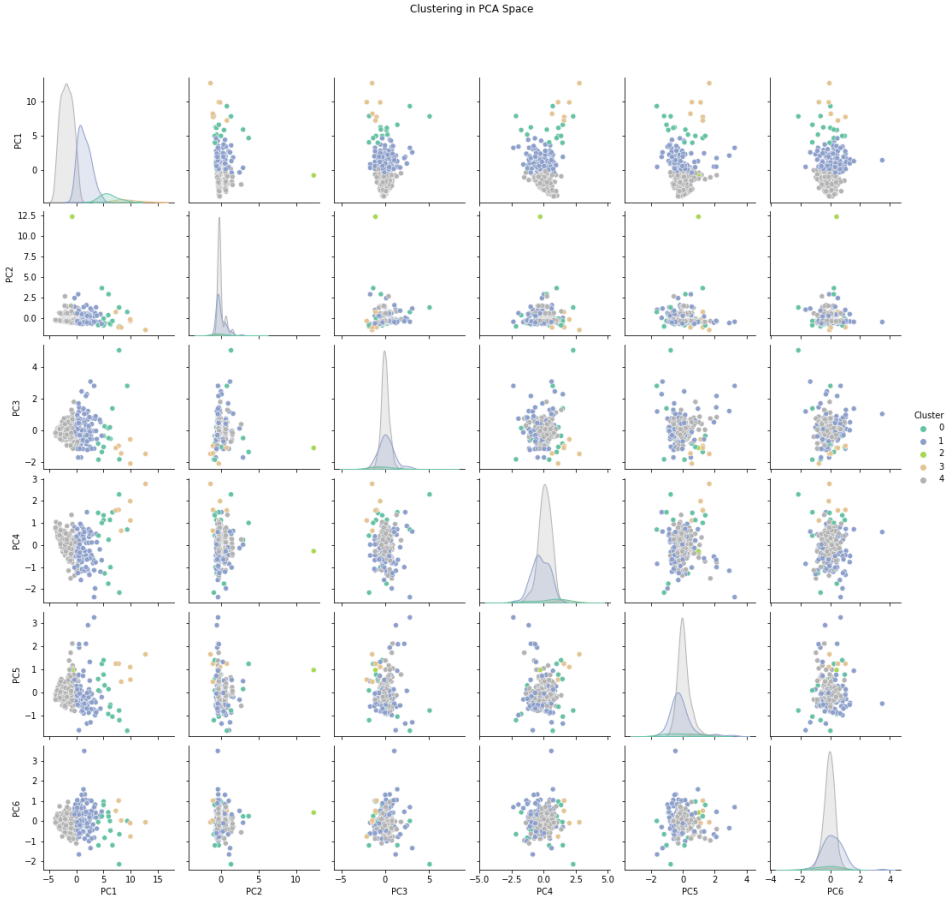


Figure A.8.1. Mean feature value per beat for each cluster in best clustering model

Mean Feature Value (count) per Police Beat by Cluster					
	Cluster				
Feature	**0**	**1**	**2**	**3**	**4**
COMPLAINTS_POLICE_SHOOTING	1	0	14	1	0
COMPLAINTS_TOTAL	52	28	30	47	11
CRIME_TOTAL	6,147	4,595	3,544	7,074	2,726
CRIME_ARREST	1,929	997	508	2,917	423
CRIME_DOMESTIC	1,307	875	534	1,226	364
ISR_TOTAL_STOPS	4,053	2,288	468	5,441	936
ISR_SEARCH	874	417	51	1,378	142
ISR_ARREST	650	287	40	1,208	102
UOF_TOTAL_COUNT	207	100	54	259	38
UOF_POLICE_W_WEAPON	14	9	4	20	3
UOF_POLICE_WO_WEAPON	67	32	12	86	11
Note: Shaded columns represent clusters with few beats assigned.					

Figure A.8.2. Proportion of feature value per cluster in best clustering model

Figure A.8.3. Proportion of cluster per feature in best clustering model

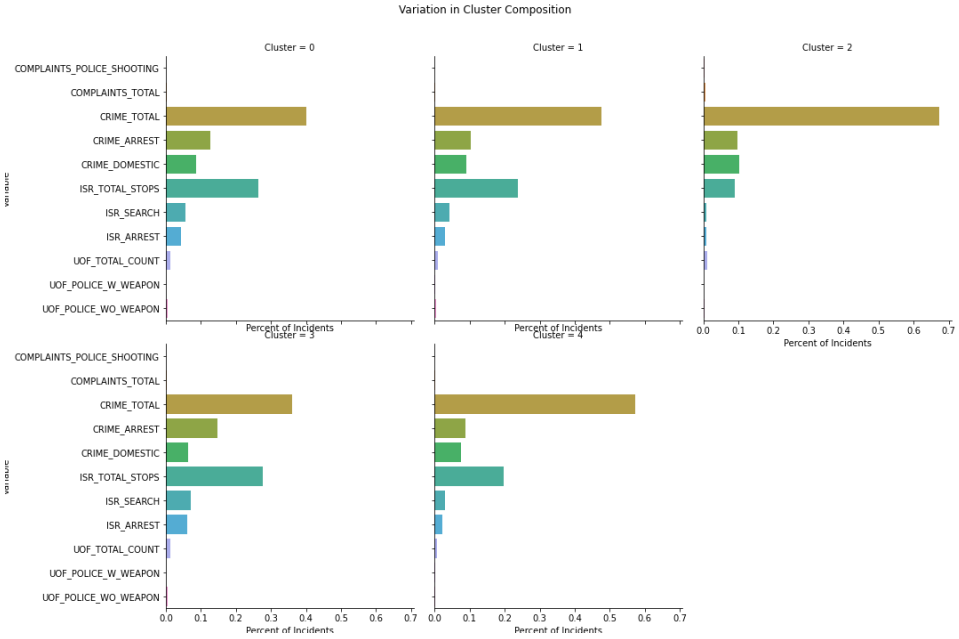


Figure A.9.1 Estimated Feature Importance from Logistic Regression

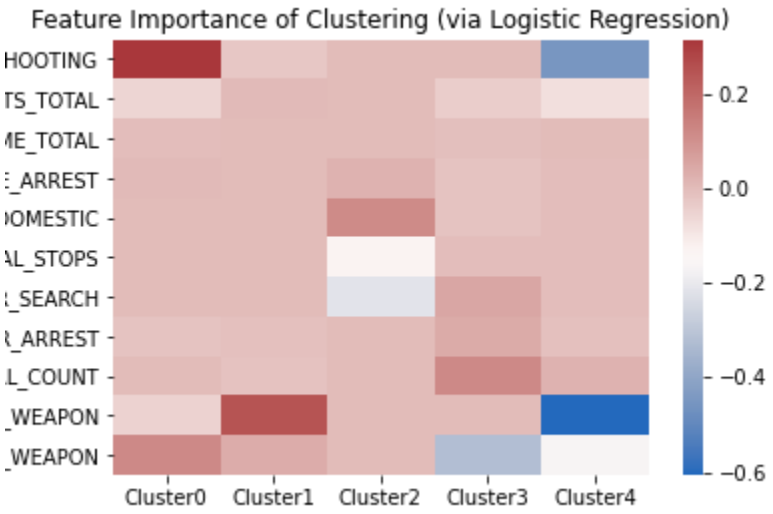


Figure A.9.2. Estimated Feature Importance from Decision Tree