Beantown Ballistics: Crime Clustering in Boston

Carolyn Fiore, Kyle Mikami, Brittany Regan
Northeastern University
440 Huntington Ave
Boston, MA 02114

Abstract

Clustering algorithms group data points together based on user-defined calculations of similarity between the points. Given accumulated data from $2015 - 2020$ that includes the racial, educational, and economic demographics of each of 12 Boston Police Department (BPD) districts as well as detailed information on every incident that BPD responded to during that timespan, this paper uses three algorithms to cluster the police districts: K-Means ++, DBSCAN, and Hierarchical Agglomerative clustering. Results are compared using the silhouette score metric. For this dataset, K-Means ++ produced the best-performing clustering solution. All associated codes and files may be found at https://github.com/insomninina/ds5230fall22/.

Introduction

Crime is always an unfortunate reality and prevalent social issue, and has recently become a key national discourse topic since the pandemic began in 2020. There are a few factors behind the increased spotlight on crime: a nationwide uptick in crime since 2019, racial disparity in policing practices, attention on police brutality and excessive force, and the accompanying Defund the Police movement. Indeed, as inflation soars there is greater focus on budgets in general, but particularly on tax dollars allocated to police forces. Recent crime statistics show the US national murder rate increased by 30% in 2020 compared to 2019, with a further 6% increase in 2021 (FBI 2022). While the city of Boston may not have seen the same spike in crime as in other major cities, there is nevertheless attention on the Boston Police Department given that it has an annual budget of over $\$400$ million, the second highest line item (CCA 2022).

With these points in mind, the goal of our project is to examine Boston crime data from a holistic standpoint. Given Boston's geography, diverse mix of neighborhoods, and its unique racial history, its neighborhoods and types of crimes by area are typically examined in silos, i.e. area by area independently. Hence, the goal of our project is to group Boston's neighborhoods based on types of crime and their prevalence. There are several questions we want to address in this analysis. First, using clustering algorithms, how do we best cluster Boston's neighborhoods into like groups based on crime types and their prevalence? Second, what is the "crime profile" of the clusters identified in terms of prevalent crime types and their frequency? Third, have these clusters of neighborhoods changed over time due to factors like gentrification? From an analysis such as this, we can not only identify neighborhoods where resources such as increased police presence can be best deployed to reduce crime, but also how different policing practices and crime mitigation strategies can be employed in various areas based on crime profiles. In other words, there should not be a "one size fits all" approach to addressing crime, particularly in a city like Boston!

Background

Clustering algorithms work to group a set of data measurements based only upon information in the data that describes the relationships between the data (TSK 2018), using a specification of similarity (e.g. cosine similarity, Euclidean distance, Jaccard similarity, etc.). The results will vary based on the algorithm and type of similarity used. Since unsupervised learning has no labels with which to verify the correctness of algorithm results, the use of multiple algorithms is common. This enables comparison of results and, in the case of time series data, a comparison of trends over time. Additionally, an evaluation metric is necessary to quantitatively compare the results of each algorithm. This research used K-Means ++, DBSCAN, and Hierarchical Agglomerative algorithms to cluster the police districts of Boston based on similarity of crime type and frequency; the silhouette score was used to assess the effectiveness of these three algorithms. For this study, only complete clustering was accepted, where each object must be assigned to a cluster. While this may allow the possibility of noise or outliers impacting the clustering solution, it enabled us to force clustering solutions that included all 12 districts.

K-Means ++ Clustering

K-Means ++ is a partitional clustering technique, which means it divides the measures of a dataset into non-overlapping clusters where each measure can only be assigned to a single cluster (TSK 2018). The basic K-Means algorithm takes an input value k and chooses k initial centroids, assigning all data points to the closest centroid, forming k clusters. Then, it iteratively updates the centroids

based on the mean of all points in each cluster until the centroids remain stable. K-Means ++ is a newer approach that chooses centroids incrementally, selecting the next centroid point by a probability inverse to each point's distance to the closest centroid. This produces a solution with lower sum of squared error (SSE) and prevents the selection of empty clusters by selecting initial centroids from the available data points. It finds the optimal clustering solution within $O(\log k)$ computing time and using $O((m + k)n)$ space, where m is the number of points and n is the number of attributes (TSK 2018).

DBSCAN Clustering

DBSCAN is a density-based clustering algorithm that calculates a probability of each measure belonging to each cluster, and assigns that measure to the cluster for which it has the highest probability (TSK 2018). The algorithm requires an epsilon value, which is the maximum distance between two points in a cluster, and a number of minimum points for each cluster, which allows it to label all points as core points (points that have at least the minimum number of neighbor points within the epsilon distance), border points (points which don't meet the criteria for core points but which are within the cluster of a core point), and nose points (the remainder of points, which are outside the cluster (TSK 2018). Careful analysis is required to determine appropriate parameters for epsilon and minimum points. The time complexity can run from $O(m \log m)$ to $O(m^2)$, where m is the number of points, but the algorithm has a very low space requirement of $O(m)$, since very little must be stored in order to run the algorithm (TSK 2018).

Hierarchical Agglomerative Clustering

Hierarchical clustering differs from the first two types discussed in that it is nested, that is, that it permits overlapping subsets within clusters (TSK 2018). The set of clusters forms a tree where the root is $n = 1$ clusters (all items in one cluster) and each cluster is the set of its downtree leaves. Continuing far enough down the tree will result in $n = k$ clusters where k is the number of items in the dataset. Visualizations of this tree are called a dendrogram, which is used to determine the number of clusters, rather than the elbow method used in partitional clustering methods like K-Means ++ and DBSCAN. For this project, the clusters were weighted, which means that the number of points in each cluster were not taken into account as part of the clustering method, and the hierarchical clustering was accomplished using Ward's method, which defines the distance between two clusters as the change in SSE resulting from combining those two clusters. Clusters whose combination would result in a significantly higher SSE are therefore considered distant. The downside to using this algorithm is that the districts are not intrinsically hierarchical, but using the similarity in crime profile as the distance between points enables appropriate use of this algorithm as a comparison to the results of K-Means ++ and DBSCAN. Hierarchical agglomerative clustering requires $O(m^2 \log m)$ computing time and $O(m^2)$ space, making it the most expensive algorithm of the three.

Silhouette Score

Silhouette score is a metric of cluster evaluation that measures the cohesion and separation of each cluster. When clustering, the ideal is for the points within a cluster to be compact, and for each cluster to be well-separated from the others. The silhouette score provides a metric for determining how well the resultant clusters score in these areas, which is informative to determine whether the clusters are distinguishable from the clusters that would form if the data were random. The value of the silhouette score ranges from -1 to 1, where -1 indicates that the inter-cluster and intra-cluster distances are not differentiable from randomness, 0 indicates the presence of overlapping clusters, and 1 indicates cohesive, well-separated clusters (TSK 2018).

Related Work

We reviewed articles of projects that clustered crimes in Canada and by US states before beginning our own. The other projects we saw focused on clustering states and cities. While this is interesting to see, we wanted to look at a more granular level since the data will vary so much within a state. The most common crimes in cities for example are likely to be very different from rural areas. Because of this, we chose to look at police districts in Boston. The other analyses also tended to use very limited measures of $3-5$ crimes or crime groups, and none conducted time series analysis. We were more interested in keeping the detailed incident level provided by the Boston Police Department (BPD) dataset. In this data, there were 54 semi-summarized crime types as opposed to a handful like in other analyses we researched. Additionally, we added census data to our project which the other projects did not include. This gave us additional information for the clusters as well as the ability to compare any changes in clusters over time to changes in census data. This enabled our clustering to include changes of the demographics and socio-economic status of each police district over time as well as simply clustering the incidents that merited a police response.

Project Description

The first steps for our project involved cleaning and combining our datasets. The crime description we wanted to use was not populated for all rows, so we used historical matches to fill in the missing fields based on other crime columns. For the census data, the median age by neighborhood was missing for the most recent year. Since the census data was slowly shifting for some neighborhoods, it made the most sense to fill the null values with the median age from the most recent year. The census data also needed to be aggregated to the police district level in order to be joined to the crime dataset. This was done by taking a weighted average of each field based on the population in the neighborhood.

Once the datasets were joined together, we conducted some initial exploratory analysis. Most of what we found was in line with what we would expect to see, but it was still good to confirm our understanding of the datasets. As a few examples, we could see that the most frequent crime in Boston was motor vehicle accidents and we could see

that while some crimes such as drug violations occurred fairly evenly across all neighborhoods, others such as homicides were clustered closer together. With the census data we could also see trends that we knew occurred. For example we can see a drop in median age in East Boston at the same time as a drop in poverty rate which corresponds to the influx of younger working professionals into the newer, more expensive housing that has been built.

The first steps for our project involved cleaning and combining our datasets. The crime description we wanted to use was not populated for all rows, so we used historical matches to fill in the missing fields based on other crime columns. For the census data, the median age by neighborhood was missing for the most recent year. Since the census data was slowly shifting for some neighborhoods, it made the most sense to fill the null values with the median age from the most recent year. The census data also needed to be aggregated to the police district level in order to be joined to the crime dataset. This was done by taking a weighted average of each field based on the population in the neighborhood.

Once the datasets were joined together, we conducted some initial exploratory analysis. Most of what we found was in line with what we would expect to see, but it was still good to confirm our understanding of the datasets (Fig. $1-2$).

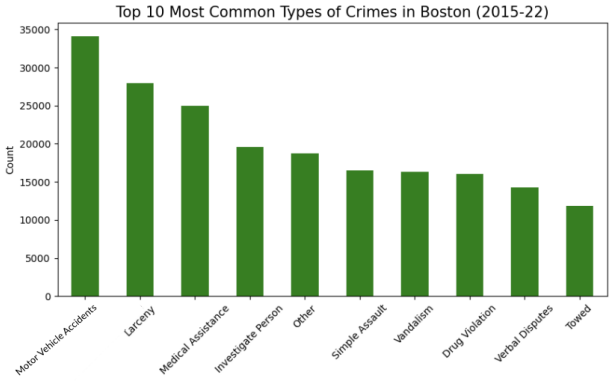

Figure 1: 10 Most Common Crime Types

A few examples: we could see that the most frequent crime in Boston was motor vehicle accidents and we could see that while some crimes such as drug violations occurred fairly evenly across all neighborhoods, others such as homicides were clustered closer together. With the census data we could also see trends that we knew occurred. For example in Fig. $3-4$, we can see a drop in median age in East Boston at the same time as a drop in poverty rate which corresponds to the influx of younger working professionals into the newer, more expensive housing that has been built.

Next it was time to try different clustering algorithms on our data. We decided to use KMeans++, DBSCAN, and Hierarchical Clustering to determine which algorithm produced the best results. For each algorithm we selected the parameters that were optimal using all of the years combined, even if they resulted in a different number of clus-

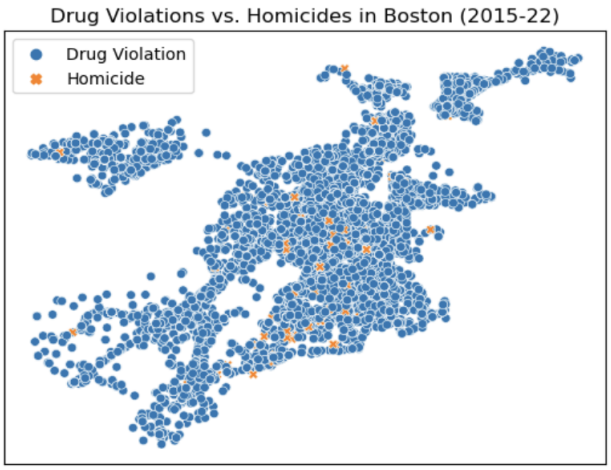

Figure 2: Drug Violation and Homicide Locations

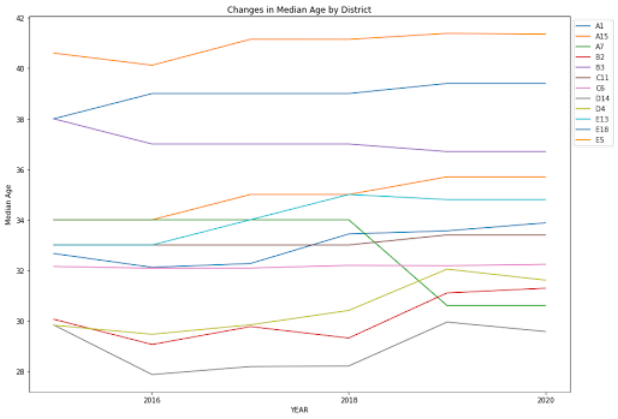

Figure 3: Median Age by District by Year

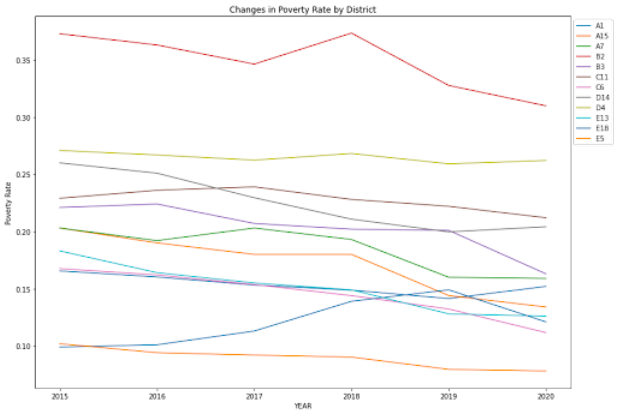

Figure 4: Poverty Rate by District by Year

ters. In order to compare each algorithm and select the best to use, we used the silhouette scores. We then ran individual years through each model to see how the clusters would change over time.

Empirical Results

For each of K-Means ++, DBSCAN, and Hierarchical clustering, we performed clustering on the dataset as a whole as well as by-year time series analysis. At the conclusion of the clustering, we calculated the silhouette scores and compared the different clustering solutions against six of the crime type category variables: assault, verbal disputes, larceny, drug violation, vandalism, and investigation of a person. The specific incidents were grouped by cluster and calculated per $1,000$ people based on the population of each district. We chose to compare two each from violent crimes, non-violent police intervention incidents, and administrative incidents to compare how the different clustering solutions affected these categories.

K-Means ++

We experimented with and optimized both the number of clusters and the number of iterations run in the algorithm. Testing the number of clusters from 1 to the number of districts and then plotting the inertia of the algorithm as a function of the number of clusters showed that the elbow was around 4 clusters (Fig. 5). Different numbers of iterations did not significantly affect the results, so we used 10 iterations.

Running K-Means ++ with 4 clusters on the overall dataset from $2015 - 2020$ and across each individual year in that timeframe yielded interesting results (Fig. 6). There are a few key takeaways. For one, Mattapan is always its own cluster, both in the overall results and in each year, indicating that it has its own distinct crime profile. Second, the downtown areas such as Downtown, North Station, and Beacon Hill, typically comprise their own cluster as well, but in some years are grouped with the South End/Back Bay/Fenway area. This makes sense given the similarities between the two areas. Interestingly, South Boston was clustered with those areas in 2020, perhaps suggesting that Southie is becoming more like downtown in terms of crime as it becomes even more urbanized and commercial. Third, the Roxbury/Mission Hill neighborhood is clustered with areas like Hyde Park, Jamaica Plain, and Southie. Finally, many of the neighborhoods around the perimeter of the city, such as Charlestown, East Boston, Dorchester, and Allston/Brighton, are typically clustered together. In some years, this cluster also includes Jamaica Plain and Hyde Park. This result makes sense again given that these areas comprise the more residential areas of Boston. They also comprise most of the land area of the city.

When we examine specific key types of crimes per capita (per $1,000$ people) at the cluster level for the clusters identified with K-Means, there are still further noteworthy results (Fig. 7). As expected, the Mattapan cluster over-indexes in most types of crimes. On the opposite end of the spectrum, Cluster 0, which comprises the more residential neighborhoods, has significantly lower rates of crimes across all types

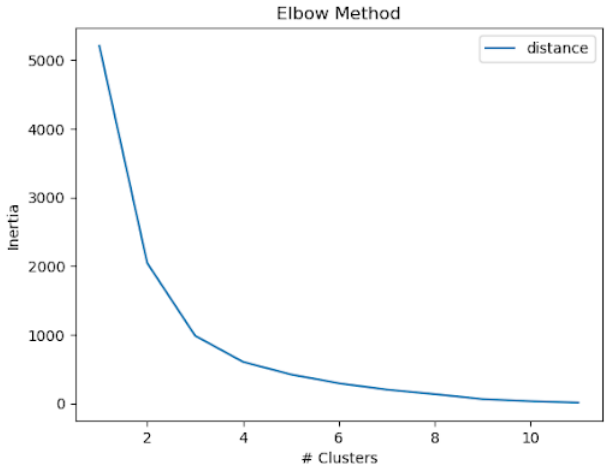

Figure 5: Elbow Curve for K-Means ++

Out[27]:

	DISTRICT	Neighborhoods	Overall Cluster	2015	2016	2017	2018	2019	2020
0	A1	Downtown areas	3	2	2	0	3	2	2
1	A15	Charlestown	0	1	0	3	0	0	0
2	A7	East Boston	0	1	0	3	0	0	0
3	B2	Roxbury/Longwood/Mission Hill	1	0	3	2	2	3	3
4	B3	Mattapan	2	3	1	1	1	1	1
5	C11	Dorchester	0	1	0	3	0	0	0
6	C6	South Boston	1	0	3	2	2	3	2
7	D14	Allston/Brighton	0	1	0	3	0	0	0
8	D4	South End/Back Bay/Fenway	3	0	2	0	2	3	2
9	E13	Jamaica Plain	1	1	3	2	0	3	0
10	E18	Hyde Park	1	1	3	2	0	3	3
11	E5	West Roxbury/Roslindale	0	1	0	3	0	0	0

Figure 6: K-Means ++ Clustering Results

Select Crimes Per 1,000 People

Cluster #	Neighborhoods	Assault	Verbal Disputes	Investigate Person	Larceny	Vandalism	Drug Violation
0	Charlestown, East Boston, Dorchester Allston/Brighton, West Roxbury/Roslindale	4.93	2.82	5.31	8.10	4.11	3.17
1	Roxbury/Longwood/ Mission Hill, South Boston, Jamaica Plain, Hyde Park	9.59	5.63	8.66	14.51	6.46	6.66
2	Mattapan	29.77	26.81	26.67	22.59	19.25	14.37
3	Downtown areas, South End/Back Bay/Fenway	13.72	1.88	7.81	31.67	6.10	7.32

Figure 7: Crime Profiles of Clusters Identified by K-Means++

compared to the other clusters. For instance, assault is 2x more prevalent in Cluster 1 compared to Cluster 0, and 3x in Cluster 3 compared to Cluster 0. In addition, as may be expected, the downtown cluster has higher larceny rates than the other clusters, and is higher than the other clusters (except Mattapan), in terms of drug violations, robbery, assault, and lost property.

DBSCAN

Before performing the DBSCAN algorithm, the elbow method was used to determine the value for epsilon. This method computes the distance to the 1st nearest neighbor for each point. One was selected as the minimum points since we wanted every district to be included in a cluster, and clusters of one were allowed. This gave a value of 9, which was used with a min-points value of 1 (Fig. 8).

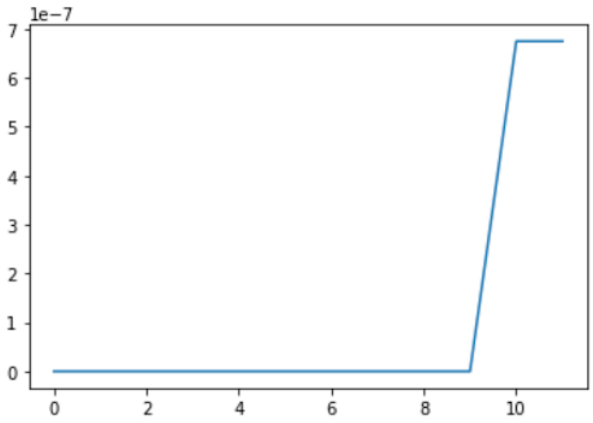

Figure 8: Selection of Epsilon for DBSCAN Algorithm

	DISTRICT	All_Years	2015	2016	2017	2018	2019	2020
0	A1	0	0	0	0	0	0	0
1	A15	1	1	1	1	1	1	1
2	A7	1	1	2	1	1	1	1
3	B2	2	1	3	2	2	2	2
4	B3	3	2	4	3	3	3	3
5	C11	1	1	2	1	1	1	1
6	C6	4	1	5	4	4	4	4
7	D14	1	1	6	1	1	1	1
8	D4	5	3	7	5	5	5	0
9	E13	1	1	1	1	1	1	1
10	E18	1	1	8	6	6	6	5
11	E5	1	1	2	1	1	7	6

Figure 9: DBSCAN Clustering Results

Using 9 as the epsilon yielded 8 clusters for each year, which is much higher than the other algorithms (Fig. 9).

Select Types of Crimes (per 1K People)

Cluster #	Neighborhoods	ASSAULT	VERBAL DISPUTES	INVESTIGATE PERSON	LARCENY	VANDALISM	DRUG VIOLATION
0	Downtown, North End, West End, Chinatown, Beacon Hill	18.35	1.59	9.11	35.16	7.07	9.93
1	Charlestown, East Boston, Dorchester, Brighton, Allston, Jamaica Plain, Hyde Park, West Roxbury, Roslindale	5.56	3.44	6.09	9.30	4.43	3.67
2	Roxbury, Longwood, Mission Hill	13.92	9.09	9.52	14.60	8.01	7.70
3	Mattapan	29.77	26.81	26.67	22.59	19.25	14.37
4	South Boston, South Boston Waterfront	10.16	3.47	9.04	18.84	7.34	9.12
5	South End, Back Bay, Fenway	9.09	2.17	6.52	28.18	5.13	4.71

Figure 10: Crime Profiles of Clusters Identified by DBSCAN

Given that there wasn't much of an elbow graph and the high resulting clusters, it made more sense to choose an epsilon value of 10. In addition to aligning more closely with the other algorithms, this also resulted in a higher silhouette score. It was interesting that DBSCAN only grouped together a lot of the neighborhoods outside of the downtown area in Boston, while the rest of the districts were clusters of one district.

When looking at the clusters over time it was also interesting how much movement there was and how varied the number of clusters were. District A1, downtown, is the only district that remained as its own cluster. Districts A15 and E13 also did not change clusters, although the other districts within the cluster did vary. Then in 2016, there were 9 clusters for only 12 districts, up from 4 clusters in 2015. Given the frequency of the changes and the knowledge that crime trends would not change that often, DBSCAN does not appear to be a good clustering algorithm for our dataset.

When examining the results of crimes per $1,000$ people grouped by the DBSCAN clusters (Fig. 10), we find that the districts that were clustered together in cluster 1 tend to have a lower crime rate per capita across the chosen crimes. This is especially noticeable with larceny which is one of the highest frequency crimes. Cluster 3 appears to be the opposite, with high frequencies across all. The only crime where cluster 3 does not have the highest frequency is larceny, where both clusters 0 and 5 occur more often. Clusters 2 and 4 are fairly similar, with slight differences in which frequencies. For example, cluster 2 has more verbal disputes and 4 has more larceny and drug violations. A lot of this makes sense intuitively since we know Mattapan has higher crime rates compared to other neighborhoods and theft is more likely when you're downtown.

Hierarchical

The Hierarchical Agglomerative Clustering algorithm uses the dendrogram rather than the elbow method to optimize the clustering parameters.

The dendrogram (Fig. 11) shows a large "distance" between the crime profile of Mattapan (district B3) and the other 10 districts. The next split on the tree creates a cluster of 6 and a cluster of 4 with the same "distance" between their crime profiles as the Mattapan split. Farther down the tree the "distances" are smaller magnitude ($\frac{1}{3}$ or less of the first two branchings). Stopping at 3 clusters resulted in a higher

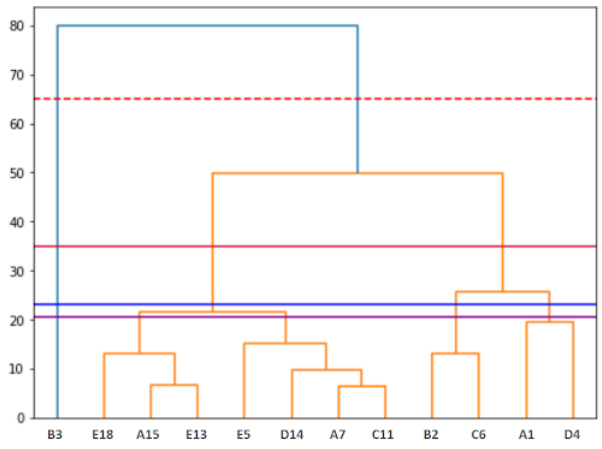

Figure 11: Hierarchical Agglomerative Dendrogram

	DISTRICT	Neighborhoods	2015	2016	2017	2018	2019	2020	Overall
0	A1	Downtown areas	3	0	4	4	2	1	0
1	A15	Charlestown	0	2	2	2	1	2	4
2	A7	East Boston	0	1	2	2	1	2	2
3	B2	Roxbury/Longwood/Mission Hill	2	4	0	0	0	0	1
4	B3	Mattapan	4	3	3	3	3	3	3
5	C11	Dorchester	0	1	2	2	1	2	2
6	C6	South Boston	2	2	0	0	0	0	1
7	D14	Allston/Brighton	0	1	2	2	1	2	2
8	D4	South End/Back Bay/Fenway	1	0	1	1	4	1	0
9	E13	Jamaica Plain	0	2	2	0	1	2	4
10	E18	Hyde Park	0	2	0	0	0	0	4
11	E5	West Roxbury/Roslindale	0	1	2	2	1	4	2

Figure 12: Hierarchical Clustering Results

Select Crimes Per 1,000 People

Cluster #	Neighborhoods	ASSAULT	VERBAL DISPUTES	INVESTIGATE PERSON	LARCENY	VANDALISM	DRUG VIOLATION
0	Downtown Areas, South End/Back Bay/Fenway	13.7181	1.8805	7.8131	31.6695	6.0991	7.3176
1	Roxbury/Longwood/ Mission Hill, South Boston	12.0422	6.2800	9.2796	16.7232	7.6770	8.4076
2	East Boston, Dorchester, Allston/Brighton, West Roxbury/Roslindale	4.7599	2.8884	4.9619	7.1558	3.8737	2.9724
3	Mattapan	29.7727	26.8071	26.6725	22.5941	19.2476	14.3683
4	Charlestown, Jamaica Plain, Hyde Park	6.6319	4.1729	7.6044	12.1538	5.1770	4.5887

Figure 13: Crime Profiles of Clusters Identified by Hierarchical Agglomerative Clustering

silhouette score (0.606 compared to 0.203) but was not comparable to the results of the other two clustering methods for analysis, so the chosen tree depth is that which split the districts into 5 clusters of sizes$\{1, 3, 4, 2, 2\}$ (Fig. 12). This did produce comparable results to the other two algorithms. Overall and for each year, Mattapan is always in its own single-district cluster, Downtown Areas tended to join with Back Bay/Fenway if not in a single-district cluster, and the peripheral districts (Charlestown, East Boston, Dorchester, and Allston/Brighton) are typically clustered together in the Hierarchical algorithms as well. And, as in the other two algorithms, the disparities between the residential areas are relatively low.

When examining the clusters' prevalent crime rates per capita (Fig. 13), Mattapan is again over-indexed on most of the crime types, although The one outlier crime type is larceny, where instead of Mattapan outstripping the rest, cluster 0, which consists of Downtown Areas, South End/Back Bay/Fenway (31 per capita), is over 4x as high as the lowest incidence in cluster 2 (7 per capita).Unlike K-Means ++, the Hierarchical clustering did not show a shift for South Boston from clustering with Roxbury/Longwood /Mission Hill to the Downtown or South End areas. The D4 district (South End/Back Bay/Fenway) was also a single-district cluster for 4 out of the six years used in this project, so its crime profile, like Mattapan's, is more unique than it is similar to the other 9 districts.

Evaluation

In addition to exploring the similarities between district groupings, we used the silhouette scores of the resulting algorithms to determine the rankings of each algorithm's clusters.

```
Overall:  0.2923654158552943
2015:  0.34360075520943784
2016:  0.2637844600236229
2017:  0.2531092504973347
2018:  0.27452847596946783
2019:  0.2595272263301555
2020:  0.3318716709427133
```

Figure 14: Silhouette Scores: K-Means ++

Based on the silhouette scores (Fig. $14 - 16$), K-Means ++ produced the best clusters in terms of cohesiveness and separation. Hierarchical agglomerative clustering produced the second-highest silhouette score based on these comparisons, but did have a higher silhouette score when selecting 3 clusters of size $\{1, 6, 4\}$. DBSCAN had the lowest silhouette scores, and the only algorithm that produced any negative silhouette score. This is likely because of the way that we used the crime profiles as similarity metrics, which may not have provided appropriate indications of density in the char-

```
Overall:  0.2923654158552943
2015:  0.34360075520943784
2016:  0.2637844600236229
2017:  0.2531092504973347
2018:  0.27452847596946783
2019:  0.2595272263301555
2020:  0.3318716709427133
```

Figure 15: Silhouette Scores: DBSCAN

```
Overall:  0.2923654158552943
2015:  0.34360075520943784
2016:  0.2637844600236229
2017:  0.2531092504973347
2018:  0.27452847596946783
2019:  0.2595272263301555
2020:  0.3318716709427133
```

Figure 16: Silhouette Scores: Hierarchical

acteristics of each district. It is also likely that the high number of clusters appeared more similar to randomness than the results of the other two algorithms.

Conclusions/Future Directions

In summary, it was possible to cluster Boston's police districts based on types of crime and their prevalence. The three clustering algorithms performed, K-Means++, DBSCAN, and Hierarchical Agglomerative, identified different numbers of clusters and grouped neighborhoods in different ways, with varying silhouette scores. As stated previously, K-Means++ seemed to produce the best clusters in terms of cluster cohesiveness and separation. Despite the differences in the three algorithms' results, there was some overlap between findings between the three. First, Mattapan emerges as its own cluster with a unique crime profile, and higher crime rates in general. Second, the neighborhoods lying around the perimeter of the city of Boston, such as Charlestown, Dorchester, and Allston/Brighton, tended to be clustered together. This cluster, by far the largest by land area, makes sense given the more residential characteristics of these areas and the degree of gentrification that has occurred in all of them over the past several years. Third and perhaps most importantly, based on the clusters identified we did see some differences in types of crime profiles and the most prevalent types of crime. For instance, in more urbanized and commercial areas like Downtown Boston and the Back Bay, larceny is much more common than in the other clusters. At the same time, many crimes we did not specifically assess, such as motor vehicle accidents, are ubiquitous and can occur in any neighborhood. Additionally, while for comparison purposes we used $5 - 6$ clusters, in a practical setting, the selection of a number of clusters may be more or less constrained, in which case the 3-cluster solution from Hierarchical Agglomerative - or even a specifically constrained number or size of clusters - may be required. Certain resources may only be able to cover $3 - 4$ police districts, limiting cluster size; or based on the different crime profile, a single-district cluster like Mattapan may require more resources than a six-district cluster.

There are a few items we would like to examine in the course of our continued studies. First and foremost, we would like to conduct a deep dive into the demographics and demographic changes of the clusters identified by the algorithms. Boston has a unique history of segregation and "redlining" that would be valuable to more granularly explore with the Harvard Redlining dataset (Nelson 2022) and some of the equality measures available in the Census data. Second, we originally planned to incorporate more years of the BPD incident report data into the analysis. While we were able to obtain data dating back to 2012, we were unable to leverage the $2012 - 2014$ data; due to a change in the BPD reporting system, there were significant differences in how crimes were captured prior to 2015 and the data were not cross- compatible even where missing data were not an issue. In addition, it would be valuable to try to obtain historical crime data dating back even further, preferably to the 1980s or 1990s when crime was more prevalent in Boston. Lastly, the Census data is quite specific; the measures that are available are heavily stratified (i.e. by age, by gender, by racial demographic) and available down to the census block level. In the future, we would like to use the 'gdal' or 'geopandas' package to code each incident in the BPD dataset to a census block to identify trends and clusters within each police district, not simply of the district as a whole. This would enable focused police resources and preventative measures as well as indicate areas in which special training like de-escalation or mental health crisis response can be focused for maximum effect. This additional project analysis would be valuable and insightful for the city with more data and a deep dive into how crime/neighborhood clusters have changed over a longer time period.

With these considerations in mind, we would thus advise future DS 5230 students to cast a wide net for additional crime data sources and spend additional time upfront trying to standardize and consolidate historical data into the same format. This can also be a consideration moving forward in a future course for this group!

References

2020 Census Block Groups in Boston. *Analyze Boston*. https://data.boston.gov/dataset/census-2020-block-groups/.

Census Tract Data (2020). *Analyze Boston*. https://data.boston.gov/dataset/2020-census-for-boston/.

Chambers, L., with Crockford, K., and Ahmad, F. [CCA]. *Unpacking the Boston Police Budget*. ACLU Massachusetts: https://data.aclum.org/2020/06/05/unpacking-the-boston-police-budget/.

Crime in Boston (2015-2022). *Analyze Boston*. https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system/.

National Crime Statistics (2022). *Federal Bureau of Investigation Crime Data Explorer*. https://crime-data-explorer.fr.cloud.gov/pages/explorer/crime/crime -trend/.

Nelson, R., Winling, L., Marciano, R., and Connolly, N.D.B. 2022. *Redlining in Boston*. Harvard Dataverse: https://doi.org/10.7910/DVN/WXZ1XK/.

P. Tan, Steinbach, M., and Kumar, V. [TSK] eds. 2018 *Introduction to Data Mining*. Pearson: New York, NY.

Poverty Status In The Past 12 Months By Sex By Age. *American Community Survey, U.S. Census Bureau*. https://data.census.gov/cedsci/table?q=United%20States&t=Income%20and%20Poverty%3A/.