Beantown Ballistics: Clustering Boston’s Neighborhoods by Crime Types & Prevalence

PROJECT PROPOSAL, DS 5230 FALL 2022

PROJECT TEAM

Team Members: Carolyn Fiore, Kyle Mikami, Brittany Regan

We are working collaboratively on the project by sharing files and meeting up as a group. We will break down specific action tasks to split up the work; for example, we split up researching possible projects and searching for relevant data sets when we aligned on the topic. We will likely split up the tidying and preprocessing, and since we are planning on running multiple clustering algorithms on our tidy dataset, we will likely each take 1-2 algorithms to completion. We each have experience writing reports and creating presentations for work, so we are comfortable collaborating in a more ad hoc manner for product creation. This approach will ensure that we take advantage of our strengths and maximize our efficiency and thoroughness throughout the duration of the project.

PROBLEM DESCRIPTION

Crime is currently a prevalent social issue and discourse topic given the national uptick in violent crime over the past few years, particularly during 2020, the first year of the pandemic. Recent crime statistics show that the US national murder rate increased by ~30% in 2020 compared to 2019, with a further 6% increase in 2021. While our home city of Boston may not have experienced the same spike as other major US cities, it is still beneficial to examine current crime patterns in order to mitigate future crime. The goal of our project is to examine Boston crime data and attempt to group its neighborhoods into clusters based on crime types and prevalence.

There are a few key questions we want to address. First, how can we best cluster Boston’s neighborhoods into like groups based on crime factors? Second, based on the clusters found, do we see all types of crime occurring together in the same neighborhoods, or do certain areas have specific issues? Third, how have these clusters changed over time with factors like gentrification? Finally, we would also like to examine how the clusters found overlap with neighborhoods that were historically “redlined” in the 1930s-1940s. Despite being a northern, liberal city, Boston has a unique history of racism and segregation, and it is worth examining the modern effects of past policies. Further, from an analysis such as this, we can identify neighborhoods where resources such as increased police presence, community programs, focused deterrence, etc. can be best deployed to reduce crime. This question is particularly interesting now given the current political dialogue around police policies and funding.

We believe these questions can be answered computationally using publicly-available datasets detailing Boston crimes over the past several years (2015-2022). Using the raw tabular data captured in these reports, including offense description, date/time, and location inputs, we are planning to use a variety of clustering algorithms and techniques in order to group neighborhoods together based on types of crime and their frequencies. A rich amount of other publicly-available data also exists which we can layer into this analysis. The ultimate output will include maps/visuals of neighborhoods found by crime clusters, an examination of the types of crimes in certain clusters, and an assessment of how clusters may have changed over time. We are eager to get started exploring this topic!

ALGORITHMS

We expect to use a variety of clustering algorithms such as K-Means, Spectral Clustering, and Agglomerative clustering. These algorithms are appropriate since we are looking to cluster the neighborhoods based on different attributes prior to conducting additional analysis and we do not know what the shapes of our clusters look like yet. At a high level, K-Means works well if the clusters are well-separated, Spectral Clustering can find non-convex clusters, and Agglomerative can find non-globular clusters. If none of these work well for our dataset, we may try out additional clustering algorithms. Many analyses exist today that look at crime data given its importance and impact. Some use clustering to link crimes based on the offender and others use clustering to predict new crimes.

DATA SETS

We are exploring eight datasets covering crime in Boston, each covering one year from 2015-2022. These datasets are available from Analyze Boston and include all crime incident reports from its respective year, which are provided by the Boston Police Department (BPD) for all events to which officers respond. Based on the preprocessing, we may determine that 2022 data is not robust enough to be compared to previous years due to the ongoing nature of data collection. We would like to compare the results of clustering algorithms over time to determine significant pattern shifts. If we are able to identify geographic clusters of certain types of crime, we would like to use the Harvard Dataverse dataset “Redlining in Boston” and US Census data on poverty by census tract to determine whether the data clusters also align with poverty or redlining labels. We would also like to explore how these trends shift over the period from 2015-2022.

In order to accomplish this, we will need to preprocess all three datasets and join them into a usable format. This will require us to identify any missing values and determine the best way to account for them (imputation, additional “unknown” variable, etc). Additionally we will need to code a conversion from lat/long coordinates to census tracts for incidents in the Crime in Boston dataset. Once that is accomplished we will be able to join the relevant measures into one dataset for running the algorithms.

LIBRARIES AND TOOLS

We will use Python and Jupyter Notebooks in order to conduct our analysis. We will use many existing libraries as well such as Scikit-Learn for the clustering algorithms, Pandas to manipulate the datasets, and Matplotlib to plot our findings and help with initial data exploration.

RESULTS

The ideal outcome of this project will be to group Boston's neighborhoods into clusters based on crime types and frequency. Ideally we will find that at least one of the clustering algorithms we implement can cleanly differentiate between neighborhoods. Preferably, more than one algorithm will be successful, so that we may compare results across approaches. We will want to compare clusters found across these multiple algorithms, and provide insights as to why any differences exist.

Intuitively, we do anticipate that certain clusters will form based on common sense factors. For example, we would expect to see more affluent census tracts form a cluster for robbery (or more high-value robbery). In addition, we would also expect certain neighborhoods to form clusters based on more violent crimes. We expect that these clusters will likely overlap with neighborhoods that were historically redlined to some degree.

There is one main potential risk we anticipate with our approach, that certain clustering algorithms will not be able to cleanly cluster or separate Boston's neighborhoods. In this case, we can attempt a few mitigating strategies. First, we can improve our data preprocessing and incorporate additional data if needed. Second, we can tweak the algorithms through running numerous iterations, attempting different variations of each (K-Medoids vs. K-Means for instance), etc.

REFERENCES

- Nelson, Robert and Winling, LaDale and Marciano, Richard and Connolly, N.D.B. *Harvard Dataverse*. "Redlining in Boston" [2022]. [Online] available https://doi.org/10.7910/DVN/WXZ1XK, accessed 18 October 2022.
- "Crime in Boston" [2015-2022]. *Analyze Boston,* courtesy of Boston Police Department. https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system, accessed 10 October 2022.
- American Community Survey, U.S. Census Bureau. "Poverty Status In The Past 12 Months By Sex By Age" [2000-2021 available]. [Online]. Multiple datasets disaggregated by race/ethnicity available: https://data.census.gov/cedsci/table?q=United%20States&t=Income%20and%20Poverty%3A,accessed 20 October 2022. .
- Federal Bureau of Investigation Crime Data Explorer. "National Crime Statistics" [2022]. [Online] available https://crime-data-explorer.fr.cloud.gov/pages/explorer/crime/crime-trend, accessed 16 October 2022.