John Tokarowski
Ramon Sanchez
C591 L1

Where should we place police stations? Optimizing BPD Station Locations

Boston is the 34th most populous city in the United States, with a population of approximately 617,000. With an ever-evolving city, maintaining a safe environment is a priority for its citizens. The Boston Police Department employs over 2,000 officers across 12 stations to combat a violent crime rate of 706.8 per 100,000 people in 2015. In total there are 11 police districts in Boston and one main headquarters. In this project, we assume locating police stations closer to areas with historically high crime rates might deter criminal activity, and sought to optimize their placement. In addition, we chose to further analyze crime breakdown in Boston across months and districts based on historical data. While historical data is by no means a clear projection of crime in the future, given our current skill set and resources available we see this as the best way to project future crime locations.

In order to carry out this project we focused on finding historical crime data in Boston and the location of the current district police stations. The three main datasets used were:

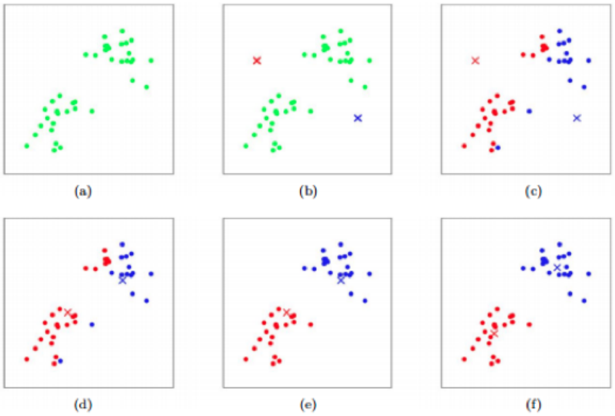
1. Boston Police Department Crime Data
2. Boston Police Department Field Investigation & Operations (FIO) Data
3. Boston Police Department Locations Data

All three datasets were obtained directly from the City of Boston Data Portal. That said, much of the data needed to be transformed in order to be used properly for our project. The main issue we faced was the need for coordinates to properly run our k-means clustering algorithm. In particular, the Boston Police Department FIO Data did not contain coordinates and instead included street addresses the officers would input into their reports. In order to fix this, we used an online API to retrieve Google map's corresponding coordinates of an address. The tool was approximately 80% accurate, and we eliminated incorrect data points located outside of Boston.

Considering that much of the data was derived from the information manually written in by police officers, we came across other obstacles when formatting the data. For one, some of the addresses included in the FIO file were nonexistent or erroneous in some other way and would need to be filtered out. Secondly, in the BPD Crime file some of the coordinates were formatted correctly but would also be found outside of Boston or even outside of the country. This was more than likely an issue caused by human error. When both the data sets were finally cleaned we merged together the FIO dataset with the Boston Crime data from the same time period and simplified the schema using a series of projections to cut down to just the necessary entries in every tuple. This allowed us to have one combined database of all crimes we could use for our k-means analysis moving forward. The final items we included in our tuples consisted of: the district, the time of day the crime took place, a description of the crime, the coordinates of the crime, the month that the crime took place in.

As stated before the the goal of this project was to analyze the distribution of crime across Boston in each month of the year and determine if police stations were optimally located to respond to crime. We specifically wanted to answer the following questions:

1. What would be the optimal location of the police stations?
2. If we were to have to close a single police district station some time throughout the year, which would be the best station to close and during what time of year?

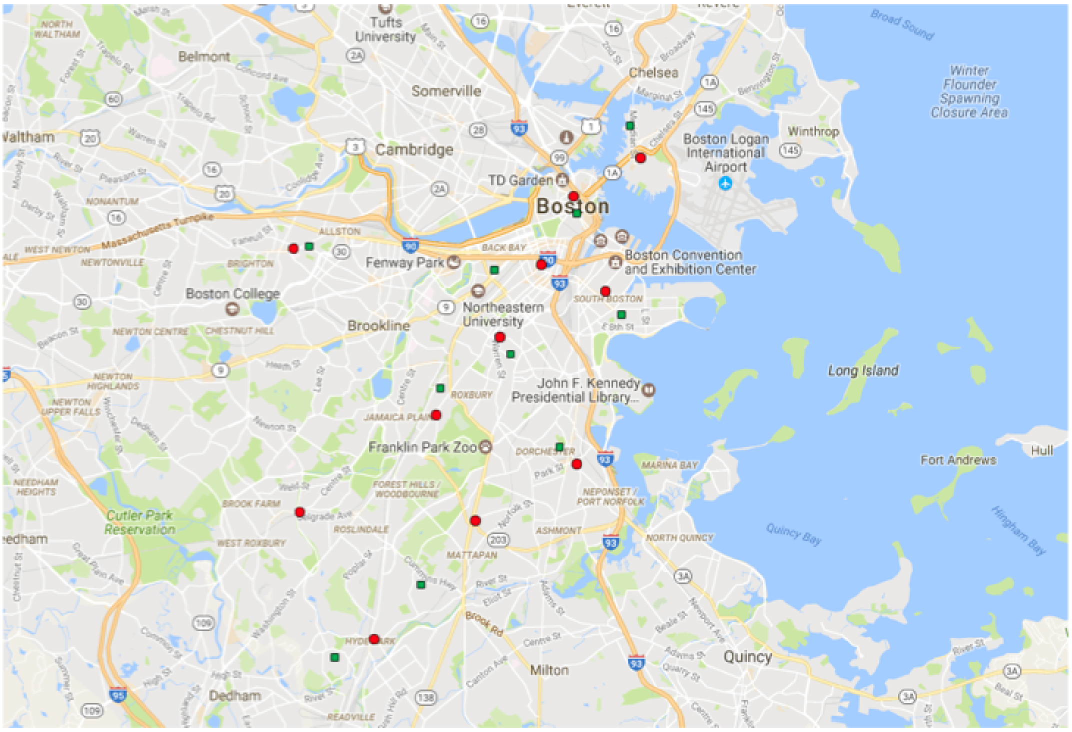


In order to solve the first question we implemented a k-means clustering algorithm using eleven means. As seen in the diagram to the right the algorithm works by calculating the distance to each random mean and assigning each point to the mean with which it has the shortest distance. As you can see the points are now organized as clusters to one of the means. Once this process is done the means are recalculated by finding the mean point of each respective cluster. Following this, the process is repeated with the new recalculated means. As the process continues you are calculating the best mean for a cluster of points. The algorithm stops once the means are no longer moving. These are now the optimal mean locations.

When implementing this process we faced multiple issues along the way. One of these issues was the size of our data set. When we originally tried to run the k-means algorithm through the entire dataset both mine and my partners computers crashed. Thus we learned very quickly we would have to run the analysis on a subset of the data until we were sure that the algorithm worked correctly. Another issue that we encountered as we implemented the k-means algorithm was that one of our means was alway (0, 0) and some of our means were unreasonable results. Originally we thought that this had to do with the original means we were passing into the algorithm so we used coordinates for the inputs that were in the center of Boston. This helped in providing us rational results. That said, we were still facing an issue with a (0, 0) mean. After trial and error we decided to print out the coordinates we were feeding the algorithm which gave us instant feedback as to what the sources of our errors were.

A standard set of coordinates in Boston would be (42.3490° N, 71.0975° W) also stated as (42.3490, -71.0975), these in fact are the coordinates of Kenmore square. The issues we were facing with our points were that some of the coordinates were missing the negative sign in the second coordinate which would result in (42.3490, 71.0975). These are the coordinates near the border of Kyrgyzstan, Kazakhstan and Uzbekistan. Another issue we found was that some of our coordinates were switched so instead of (42.3490, -71.0975) it was (-71.0975, 42.3490). These are the coordinates on the edge of Antarctica. The final point was that in some of the tuples where the coordinate data was missing, by default the coordinates (0,0) were put in place. These outliers were throwing off our results and creating optimal means far outside Boston. Once these three issues were fixed our k-means output worked.

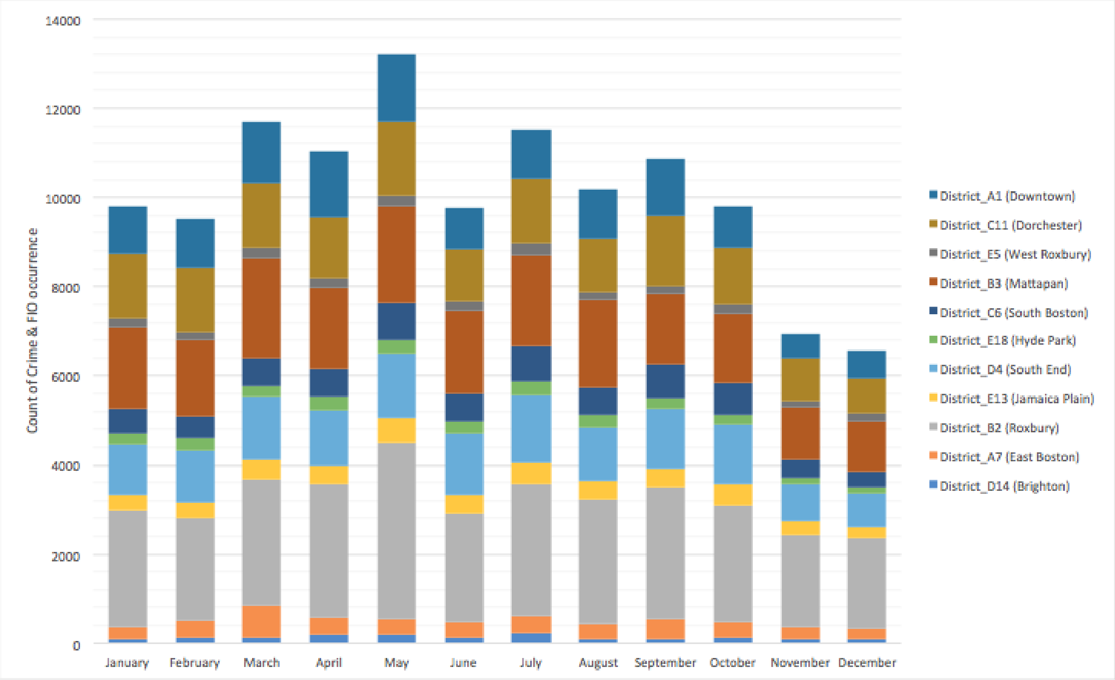
The result of this analysis is represented in the visualization below. The original police station locations are the red circles and the new optimized means are the green squares.



Running the K-means algorithm on the coordinates for every FIO and Crime occurrence coordinate moved every police station to a new location. This new location minimizes the distance to dispatch police to respond to crimes or more easily conduct field investigations and operations in high crime areas. We notice the distance of the move is less than a mile in every case. The stations in Dorchester and Mattapan barely moved, while there was a noticeably large shift in the Jamaica Plain and Roxbury stations towards the highest crime area in Roxbury. Overall, we believe the police stations are relatively well located to account for current crime occurrence, the small changes of less than a mile could be implemented but may not have a substantial impact on crime. If the analysis were to consider the traffic patterns around the new locations, it might actually take longer for the police to respond. Another explanation as to why the stations may not have moved as much may be due to the fact that police officers naturally patrol areas around their respective district station. As such, the majority of the crimes they report are situated around their police stations possibly overstating the number of crimes versus those in other areas. In addition, police stations are generally located in population centers, with more people around that area you would expect more crimes to be conducted around the center simply due to the larger amount of people.

To answer the second question, we bucketed each occurrence by district and month by using a series of selections and projections on the combined data set. For any given observation it would fall into one district and one calendar month, to contribute to a picture of which districts see the most crime and at what time of year. The dataset included both field investigations and registered crimes logged by the Boston Police Department, since both types of occurrence would require police be dispatched from a station to respond. The headquarters was omitted from this analysis because it was determined that this station would need to remain open always in order to facilitate cooperation between other districts and effectively allocate resources. The implementation of this was fairly straightforward and we did not face issues.

The result of our analysis is shown below:



After examining the data and visualizations, it was determined that the Brighton district in January experienced the lowest volume of crime and field investigations. This might be partly due to the high concentration of students in the area who would return home for the majority of the month on vacation. Additionally, Boston is extremely cold in January and a lower volume of people out on the streets could also lower crime occurrence. With an average temperature of 29 degrees Fahrenheit, and 13 inches of snow, it makes sense that the occurrence of crime would be lower as residents remain indoors. If we were to close one district for a month to cut costs and reallocate resources, Brighton in January would be the least impactful. Resources could be deployed to Roxbury, which had more than 30 times the number of occurrences over the same period.

Moving forward, future additions to this project could use a weighted k-means when carrying out the location optimization. Throughout our implementation we weighted all crimes as equal but an interesting case could be made for weighting certain crimes more than others in the k-means optimization. For example, we could weight a homicide two or three times in comparison to robberies or grand theft autos. This poses an additional question as to what would be considered a fair/best way to weight crimes. Would it be best to weight crimes based on the number years the individual would serve if found guilty? Would you weight crimes based on where they were located or when they were committed? This is a question that is open to many perspectives that we encourage the reader to pursue.