# Optimizing Boston Police Department Location using k-means algorithm

John Tokarowski
Ramon Sanchez
CS591 L1
Boston University

## Introduction

Boston is the 34th most populous city in the United States, with a population of approximately 617,000. The Boston Police Department employed over 2,000 officers across 12 stations to combat a violent crime rate of 706.8 per 100,000 people in 2015. 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 with several variations in our analysis.

## Goals & Techniques

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:

**Question 1: What would be the optimal location of the police stations?**

Solution: In order to solve this problem we focused on implementing the k-means algorithm onto our dataset and focused on finding a total of eleven means. The k-means algorithm aims to partition the observations into k clusters in which each observation belongs to the cluster with the nearest mean.

**Question 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?**

Solution: In order to solve this problem we analyzed the data and focused on finding the amount of crimes that occurred in every month of the year for each respective district.

## Analysis & Methodology

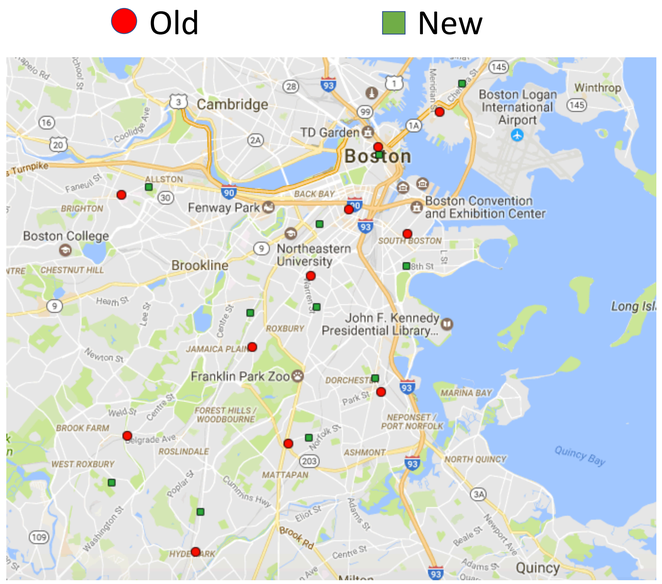
**Question 1:**

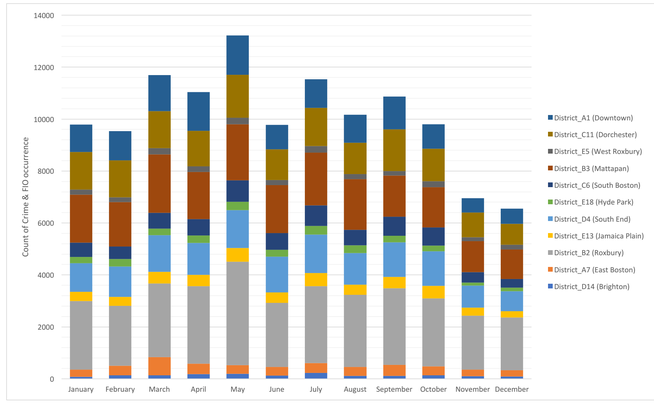
In order to solve this problem we implemented a k-means clustering algorithm using 11 means. As seen in the diagram to the right the algorithm works by calculating the distance to each random mean and assigning each point to the mean with which it has the shortest distance. As you can see the points 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 stationary and are no longer moving. These are now the optimal mean locations.

**Question 2:**

For the statistics section, 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.

## Visualizations




## Data Sets & Data Manipulation

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

The FIO dataset was limited in it's location availability. The officers entering their findings often listed a nearby address or an intersection, which could not be fed to the K-means which utilized coordinates. In order to fix this, we used an online API to retrieve Google map's best guess on coordinates of an address. The tool was approximately 80% accurate, and we eliminated incorrect data points located outside of Boston.

With the FIO data now paired with coordinates, we merged the 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 any tuple.

## Conclusion

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.

Running the K-means algorithm on the coordinates for every FIO and Crime occurrence coordinate moved every police station to a new location which 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.