Shirish Dhar

Spatial Data and Analysis - Final Project Report

Abstract

This project looks to analyze the growing nature of crimes in Berkeley to find insightful patterns in their distribution, followed by offering some corrective measures to tackle this issue using spatial analytics.

First, the Berkeley crimes distribution in 2016 is deconstructed by understanding where they occur, what time of day are they most prominent in, and what are the different types of crimes occurring within the city. Next, a more detailed analysis ensues, which locates the most crime intensive areas within Berkeley using intensity mapping, followed by the correlation of crimes with respect to the three BART stations in Berkeley, which proves to be significantly high.

Finally, the project moves into the recommendation stage, where the goal is to pinpoint the optimal location of the Berkeley Police Department based on the distribution of crimes. This is done by utilizing 'distance optimality' to all the crimes in Berkeley. Once completed, special emphasis is given to the schools within Berkeley that are most affected by crime, and the optimal location of the police department is moved closer to these schools.

Contents

Section 1 - Introduction and Motivation

Berkeley is considered safer than only 4% of the cities in the United States, being ranked the 43rd most dangerous city in America (Source - blog.sfgate.com). To make things worse, Business insider ranked UC Berkeley as the third most dangerous college in America. Over the past year, huge concern is growing amongst Berkeley students towards the influx of their inboxes with crime alerts from the Berkeley Police Department. Being a student of this wonderful university within this great city, my motivation for this project stems from this growing concern, as well as from a desire to utilize my accrued knowledge of spatial analytics to offer valuable suggestions for tackling the crimes in Berkeley.

Section 2 – Data and Preprocessing

The dataset for this project is from the City of Berkeley Open Data portal. It consists of all types of crimes within Berkeley for the year 2016. In addition to the crime locations, this dataset consists of other useful features, namely: -

- Type of Crime (Theft, Narcotics, Missing Person, Homicide, etc.)
- Date and Time of Occurrence (In the format – 'Tuesday 10/06/2016 9:40 am').

The pre-processing for this dataset was performed in Python. There were three main tasks within pre-processing: -

- Deconstructing the Date and Time column to acquire three separate columns – Event time of day, Event day of week, and Event month of year. This was done using text processing via regular expressions in Python.
- Binning the 'Type of crime' column into Violent crimes and Non-Violent crimes. This was done using data manipulation techniques in Python. Crimes like 'Homicide', 'Arson' and 'Gun/Weapon' were binned under 'Violent crimes'.
- Creating four bins for the 'Time of Day' column – Morning, Afternoon, Evening and Night. All the crimes were binned into these four categories. This was performed in Python.

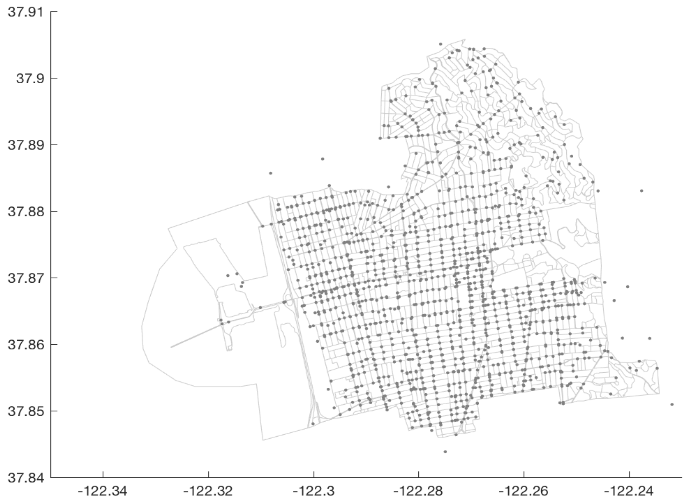
Section 3 - Analysis

The analysis for this project was divided into two parts: -

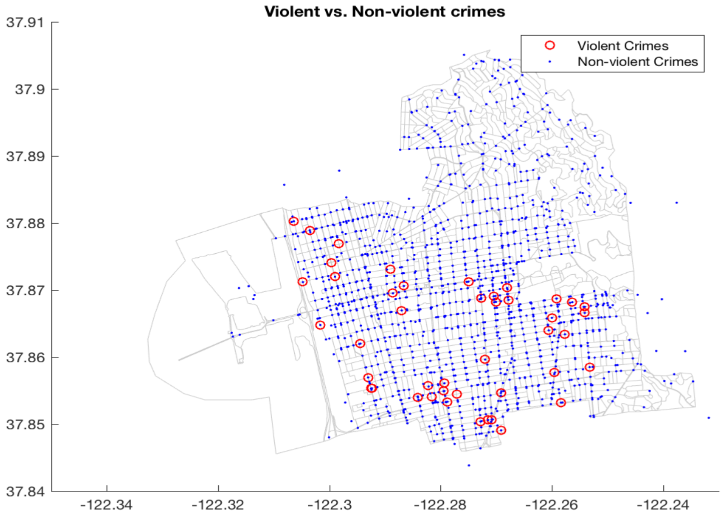
- Preliminary Analysis
- Advanced Analysis

3.1 Preliminary Analysis

The very first task in the analysis was to plot all the 2016 crimes in Berkeley and gauge if some preliminary patterns can be recognized. This is the result: -

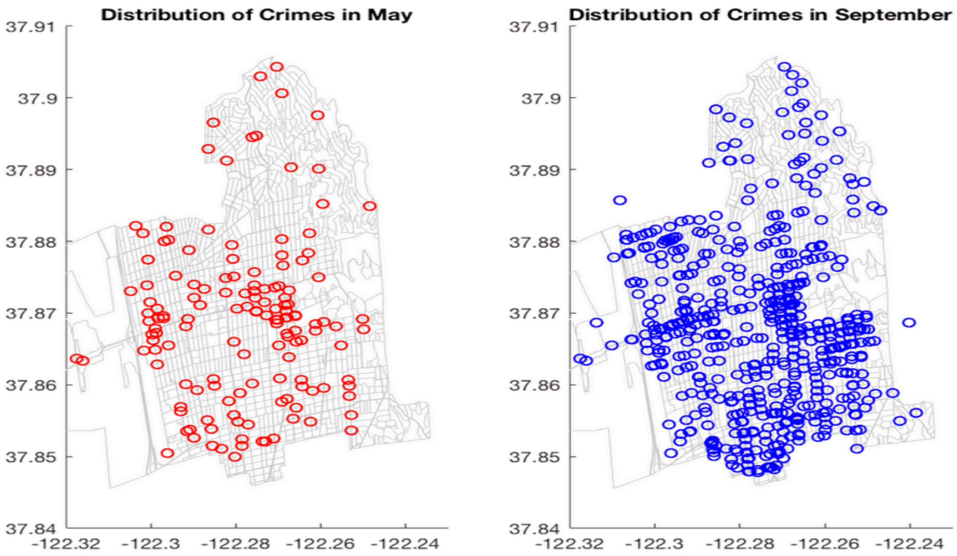


The first intriguing question was whether there was a pattern in the distribution of violent crimes in Berkeley. After binning crimes like Arson, Gun/Weapon Crimes and Homicide as 'Violent' crimes, these were highlighted to look for patterns. This is the result: -



It is very evident that all the violent crimes are located either in the southern parts of Berkeley or in the far West. There were zero violent crimes in the Northern parts of Berkeley, which is an important finding. The parts of the city just south of UC Berkeley consisted of the most number of violent crimes, which is disturbing as this is an area hugely infested with Berkeley students.

Next, it was important to find out whether there were any temporal patterns in Berkeley crimes in 2016. To find month-based patterns, the distributions of crimes for each month were plotted and compared. An example comparison between the crimes in May and September is shown: -



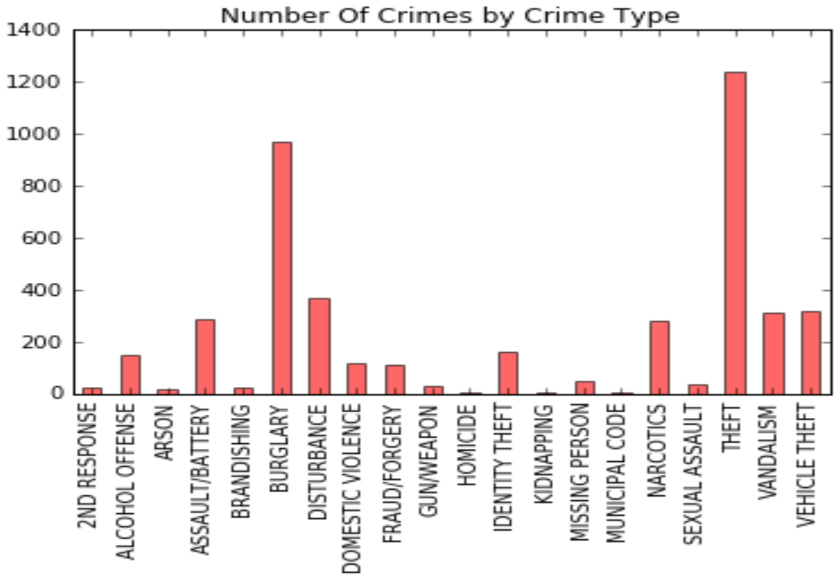
The next step was to quantitatively compare different months using numbers. An important finding that involves May and September is: -

Number of crimes in May were 167, whereas the number of crimes in September were 907.

In terms of violent crimes, May had 2, whereas September had 7.

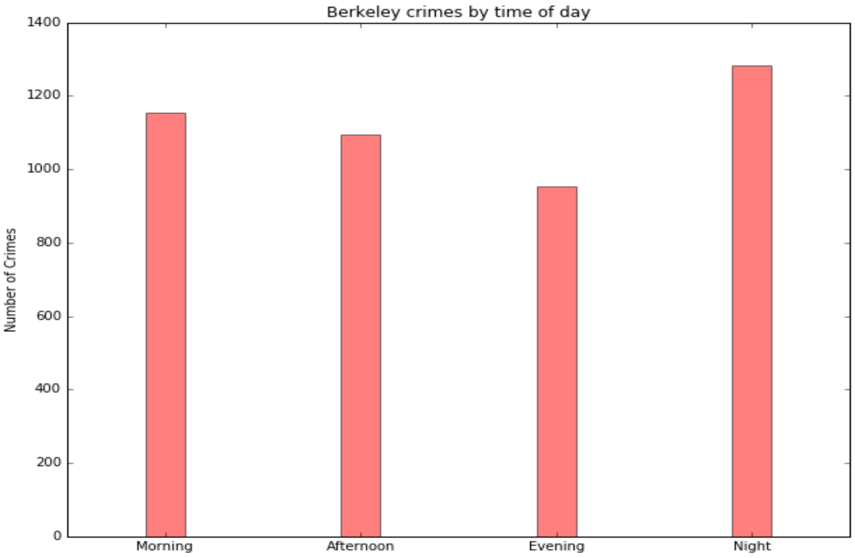
This clearly displays an insightful finding, which is that September had a much larger number of crimes in general as well as violent crimes than May.

The next step was to go deeper into the types of crimes and compare their intensities. Using matplotlib, the types of crimes in Berkeley along with their intensities were plotted.



Burglaries and thefts were the most prominent types of crimes within Berkeley in 2016, with violent crimes like arson, weapon crimes and homicides making up a very small percentage of total crimes.

The final part of the preliminary analysis was to plot the distribution of crimes based on time of day. For this, the crimes were divided into four bins – Morning, Afternoon, Evening and Night.



Some expected patterns are noticed, especially the prominence of crimes during the night.

3.2 - Advanced Analysis

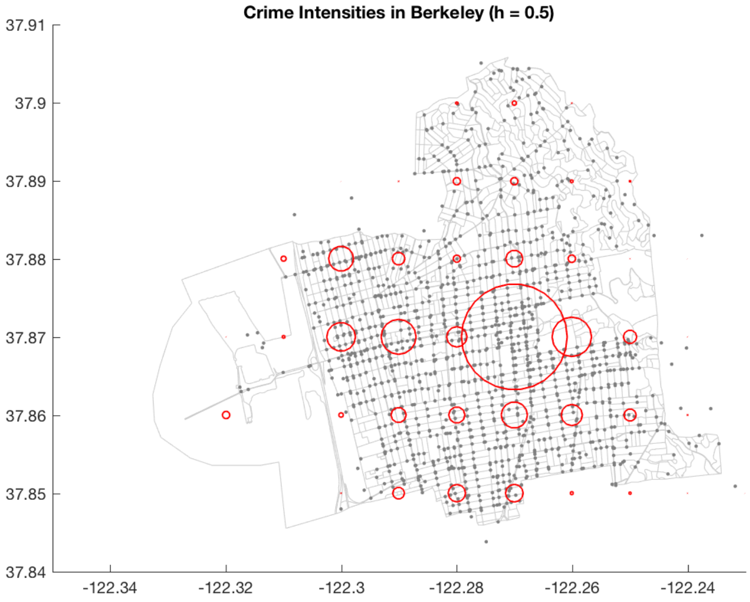
The more advanced parts of the analysis consisted of two parts: -

- Crime Intensity Mapping
- Clustering around BART Stations

3.2.1 - Crime Intensity Mapping

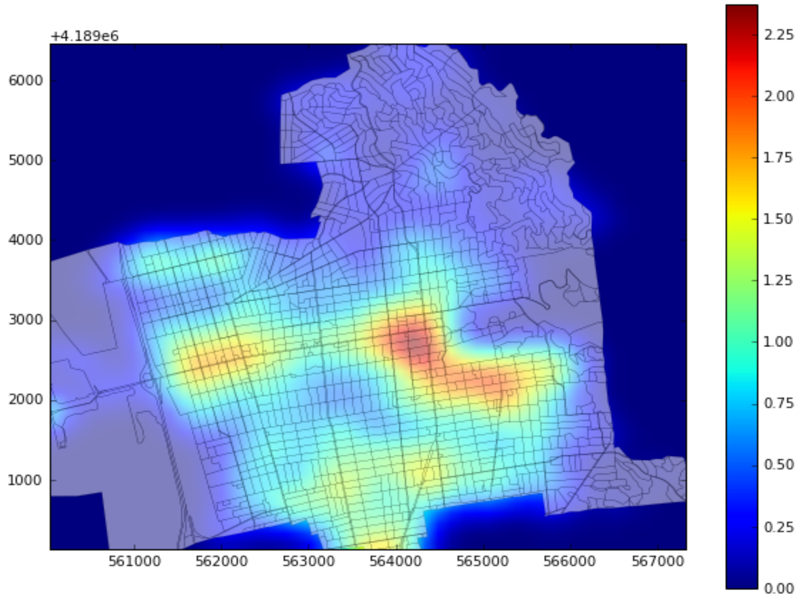
Till this point, most of the inferences around crime clusters were based on a cursory look at all the crimes in Berkeley. It was essential to find out where the biggest clustering of crimes lay within Berkeley. This was done using two ways – Intensity mapping in MATLAB and the use of heat maps in Python.

- Intensity Mapping in MATLAB



The above intensity map was created by calculating the number of crimes within a constant buffer region across Berkeley. It is evident that the area just west of the UC Berkeley campus is the most crime infested zone in Berkeley, followed by the southern parts of campus.

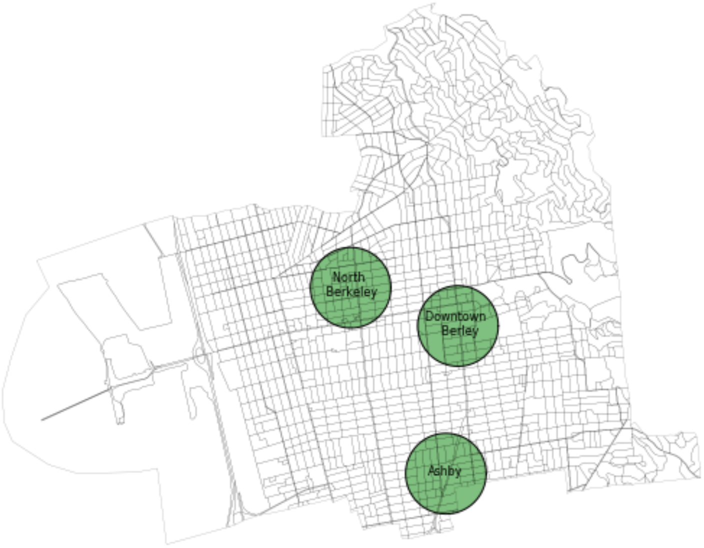
- Heat Map in Python



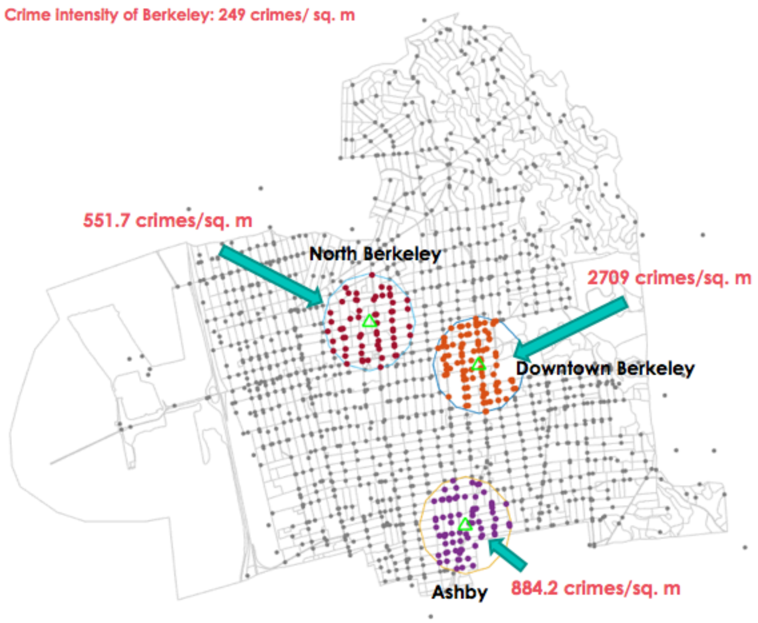
Similar insight can be gathered from this heat map. The areas just west and south of campus are the most crime intensive, and the northern part of Berkeley is relatively less crime-intensive.

3.2.2 - Clustering around BART stations

One of the most prominent hypotheses is that BART stations are epicenters of crime clusters in Berkeley. To answer the validity of this hypothesis, the BART stations in Berkeley were plotted along with a buffer of a specific distance around them.



First, the average crime intensity of Berkeley was calculated by dividing the number of crimes in 2016 by the area of Berkeley. Then, the crimes intensities around each BART station was calculated. These were then benchmarked against the average crime intensity of Berkeley. Below is a comparison: -



This shows that in general, crimes are highly clustered around the three BART stations in Berkeley, with the clustering being far above the city average of 249 crimes per square mile. Additionally, the Downtown Berkeley BART stations is a huge epicenter of crimes in Berkeley.

Section 4 – Recommendations

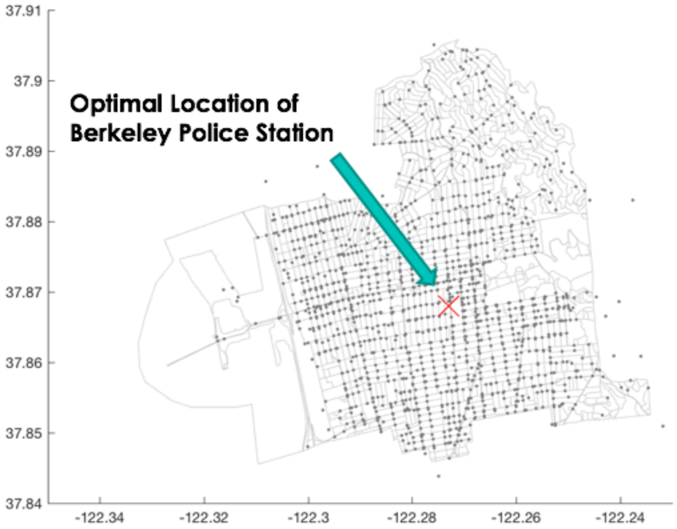
A big part of the motivation for this project was to not just analyze the growing nature of crimes in Berkeley, but to also suggest some possible remedies to tackle this growing concern. To this end, the main goal of the project is to recommend the ideal location for the Berkeley Police Department.

This was done in two parts:

- Distance Optimality to all crimes
- Distance Optimality after prioritizing crimes around schools

4.1 - Distance Optimality to all crimes

As the first step to pinpointing the optimal location of the Berkeley Police Department, a fine grid with very low spacing was superimposed on the city of Berkeley. Next, for every intersection point of the grid, the average distance to all crimes within Berkeley was calculated. Finally, the intersection point with the least average distance to all Berkeley crimes in 2016 was chosen as the optimal location for the police department.

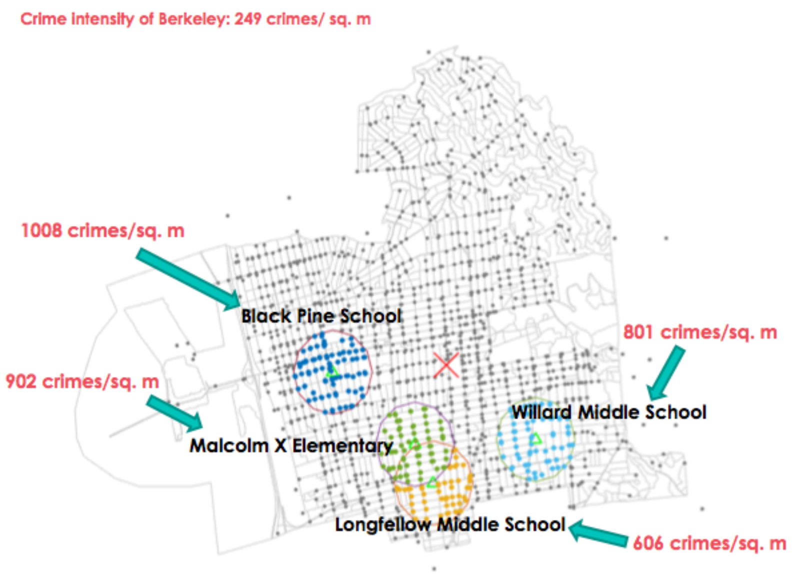


The calculated optimal location was consistent with the finding that a huge majority of the crimes was situated around the Downtown Berkeley BART station. The coordinates of this optimal police station location are (37.8680, -122.2730).

4.2 - Prioritizing crimes around schools

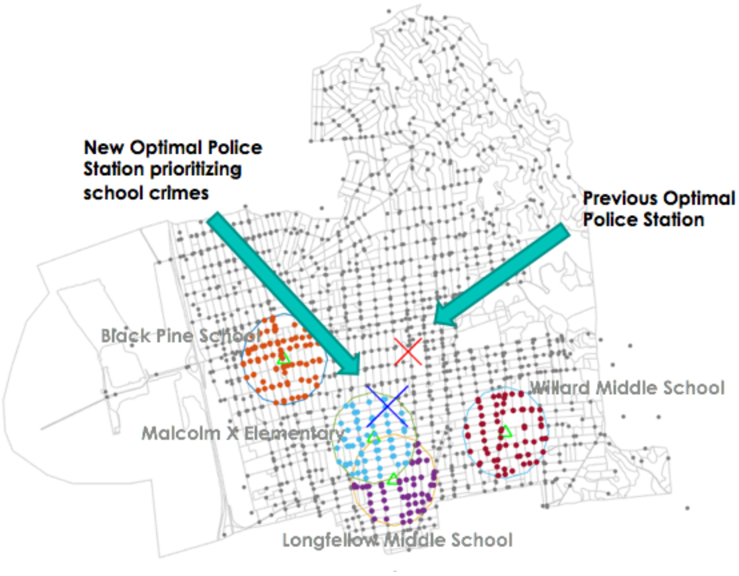
One caveat with the previous approach was that the model accounted for all crimes with the same weight/emphasis. With Berkeley becoming one of the most crime infested cities in the country, Berkeley's schools were also becoming some of the most crime-prone schools, which is extremely bad for a city known for its education. This was the primary motivation behind the decision to prioritize schools in Berkeley when calculating the optimal location for a police station.

To this end, all the schools in Berkeley were used as a starting point, and the top 4 crime-infested schools were then considered in the new model. The four most crime-affected schools in 2016 were – Black Pine School (West Berkeley), Malcolm X Elementary (South West Berkeley), Longfellow Middle School (South West Berkeley) and Willard Middle School (South of Berkeley). The crime intensities around these schools were benchmarked against the crime intensity of the city.



As is evident, these schools have a much higher crime intensity in their vicinity than the city average.

The next step was to assign a higher weight to the crimes around these four schools as compared to the other crimes when calculating the optimal location of the police department. For this, two arrays were created – Distance of grid point to non-school crimes and Distance of grid point to school crimes. When calculating the 'average distance to crimes' for each grid point, the 'distance to school crimes' array was appended five times more than the 'distance to non-school crimes' array. This meant that each school crime was given five times as much weight as all other crimes. The resulting point after prioritizing school crimes was as follows: -


New Optimal Police Station prioritizing school crimes
Previous Optimal Police Station
Black Pine School
Willard Middle School
Malcolm X Elementary
Longfellow Middle School

It is evident that the optimal location of the Berkeley Police Department moved towards the south-west direction, which is also the general direction of concentrated crime clusters in Berkeley. This new point is much closer to the crimes around the four schools, while also being optimally located to all the other crimes in Berkeley. This improved methodology allowed us to pinpoint a location that would account for all crimes in Berkeley, while also paying special attention to reducing crimes in these four crime-afflicted schools.

Section 5 - Conclusion

When analyzing the crimes in a city and attempting to recommend corrective measures for the same, many secondary factors need to be accounted for. One such factor is the safety of school children within the city. This project attempts to prioritize Berkeley's school children by placing great emphasis on the crimes around the four most crime-afflicted schools in the city. The police department in a city is pivotal to its safety, and ensuring an optimal location based on crime distribution can go a long way towards improving the safety of cities like Berkeley.

Section 6 - Challenges and Caveats

6.1 - Challenges

- While the plan was to 'prioritize' crimes around schools while calculating the optimal police station location, implementing this priority mechanism in code proved to be very difficult. Initially, the plan was to divide the distances to school crimes by a factor so that they are a fraction of their true value, and this scaled-down distance is used in further calculations. This did not truly 'prioritize' the school crimes, as it was ineffective in changing the optimal location of the police department. It was only later when a more effective way was implemented – by creating an array 'distance to all school crimes'. When calculating the average distance to crimes from each grid point, this variable was used five times as much as the array 'distance to all non-school crimes'. This way, the optimal location was shifted in the direction of school crimes.

- Pre-processing the dataset was extremely essential and difficult in this project. The occurrence of each crime was in the format "Monday 10/06/2016 9:35 am". Unpacking this into a numeric format which consisted of three columns – Time of day, Day of Week and Month of Year – was difficult. This was accomplished using natural language processing techniques like regular expressions.

6.2 - Caveats

- The biggest caveat with the recommendation is that a police department is not the only factor in tackling crime within a city. There should be equal emphasis placed on optimizing the police patrol routes, because most crimes are tackled by the closest police patrol car. Thus, optimizing the location of the Berkeley Police Department should be complimented by optimizing the police patrol routes within Berkeley, and only then will crime be tackled effectively.

- The model used in the project does not account for ‘intensity’ of crimes. Thus, crimes like homicide and arson are given equal emphasis as petty thefts when calculating the optimal police station location. A counter argument to this is that the frequency of high intensity crimes is a tiny fraction of the frequency of thefts and burglaries. By that reasoning, giving them equal emphasis makes sense.
