Haydn Kennedy and Kathryn Quirk

May 4, 2019

CS504

Final Report

Introduction

Originally unsure of where to start, we decided to browse through Boston's Open Data Portal to see if we could find a topic that piqued our interest. We found one dataset detailing the compensation for police officers in Boston. Something of note was that the compensation was much higher than we had anticipated. As a result, we became interested in the allocation and distribution of police resources within the city. Towards a solution to that problem, we've looked at both: where crimes are centered among Boston's 12 police districts, and which demographics might correlate with crime incidents across districts. With this information, we can determine which point in a given police district minimizes the mean distance to where crimes are occuring in that district and how we might allocate police resources in those districts to better address criminal activity.

We investigated three main factors' correlations from Boston census data with number of crime incidents: per capita income, education level (bachelor's degree or higher attained), and percentage of white non-hispanics in a neighborhood. For each of these factors we calculated the IQR to see if there were any outliers at the neighborhood level and also at the district level. Then, we calculated the correlation coefficient for each of the factors for each district against the number of crime incidents per district to see if there was any connection. Finally, we found the districts with the highest and lowest number of incidents and stored the details of those specific districts in our database for later reference. With these statistics, we were hoping to find anything that could help us to predict where crimes might occur.

We pulled multiple datasets, some of which we ended up using, others which we did not because they were no longer useful given the direction our project took. However, the datasets are significant as they can be used in future work. One of these

datasets was a zipped shape file from Harvard Dataverse that contained information about the police districts within Boston. In order to better use the data, we then converted it into a GeoJSON format. As a GeoJSON file, the data is more easily stored and queried from within MongoDB. The dataset itself is not too interesting, but it can be combined with some of our other datasets, e.g., the crime statistics for each district, which would allow us to better visualize crime per district.

Another dataset that we did not end up using was the salary data for all employees under the city of Boston (police, teachers, etc.). Taken from the Boston Open Data Portal, these data include total compensation, base salary, as well as overtime and various extra payments. From the data, we selected for employees of the Boston Police Department. Then, we pulled out relevant subsections of total compensation for each officer and aggregated them to compute the total amount paid for each category using MapReduce (total, overtime, detail, etc.). This transformation gave us better insight into the amounts spent towards police salaries by the City of Boston. The city spends nearly $400 million on police compensation, and there are some officers who have made upwards of $100,000 in overtime or detail work (more than their base salaries) which we find intriguing.

The last dataset that we did not use was the BPD Field Interrogation and Observation records (also from the Boston Open Data Portal). These include traffic stops, car searches, wellness checks, etc. and list officer names, dates, and incident descriptions. We initially planned to use this dataset with our police salary information to potentially look at which officers are mentioned in the most FIO entries, and by what types. However, we ended up looking at police districts rather than individual officers and so that didn't end up happening.

Of the data we did use, the neighborhood census data proved to be extremely useful for our project. The main subsets that we pulled from this dataset are the race and ethnicity make-up, the per capita income, and the level of educational achievement by neighborhoods in Boston. Due to the nature of the data (it was presented in a pdf), it was better for us to store the data in a different format and have it on the

datamechanics.io site. What is great about this data is that not only does it provide valuable insight into the demographics that make up each neighborhood of Boston, it can easily be combined with other datasets. For example, we combined it with a dataset that matched each Boston neighborhood to a police district. As a result, we are able to come up with a better picture of what the demographics are for police districts, rather than just neighborhoods. This was done using MongoDB MapReduce functionality.

And finally, we also relied on the Boston Police Department's Crime Incident Reports dataset, which documents crimes reported to and recorded by the BPD. Similarly to the FIO records mentioned above, we wanted to look at how different officers were involved in dealing with various crimes and crime trends based on location. As also mentioned earlier, we ended up only looking at crimes-by-district. These data are notably different from the FIO records, though, in that this dataset includes all activity responded to by BPD and the FIO records only include occurrences initiated by police and which only sometimes result in discovered criminal activity.

Methods

After gathering the datasets, it was important for us to decide how to make sense of them. For the census data, we realized we could add meaning to it by mapping each Boston neighborhood to the correct police district. As a result, we can have a better understanding of the demographics for a given district. Furthermore, since we had the crime incident count per district, we knew that we could then connect that data to the police district demographic data. The main motivation for combing these datasets was, as previously mentioned, if there was a good way to predict the amount of crime incidents that may occur in a police district. With this in mind, it made sense to see if there was a correlation between the census data and the crime incident data.

To see if these two data sets were connected in any way, we calculated the correlation coefficient for the following datasets: Per Capita Income vs. Crime Incidents, Race and Ethnicity vs. Crime Incidents, and Education Level vs. Crime Incidents. The per capita income dataset was straightforward to plot against the number of crime

incidents for each district because it was just the average income for each district. However, it was more difficult for the other two census datasets. The race and ethnicity dataset and the education dataset, it was broken up into categories of which we could not take an average. For example, the education dataset gave the percentage of people who had less than high school, high school graduate, some college, and Bachelor's degree or more. For our purposes, we decided to specifically look at the percentage of those with a Bachelor's degree or more for the education dataset and at the percent Non-Hispanic Whites for the race and ethnicity dataset. From here, we were then able to calculate the correlation coefficient to see if there was any correlation between the datasets mentioned. The main limitation for our data was that a linear regression did not seem to model the data well at all, using some other model may have helped to yield a better correlation coefficient.

We then wanted to see how crimes were distributed within each police district. To do so, we imported all of the crime data and cleaned it to obtain the latitude, longitude, and police district in useable formats. We then sorted the data into buckets by district and ran K-means to find the 12 points in Boston that minimize the distances between crimes. These locations were then written to .kml files along with the coordinates for Boston's police stations so that they can be visually compared.

Results

Overall, we found that there was not a strong correlation between any of the census data and the number of crime incidents per district. To make sure that there were no outliers that were affecting the correlation coefficient we even calculated the IQR for each subset of the census data. We found that there were no outliers that could have affected the data. This means that the data that we have collected cannot predict the number of crime incidents alone. One potential explanation for why we did not see as strong as a correlation as we had expected could be that we took the average of per capita income per neighborhood. However, typically in urban areas there are a lot of people on the lower and higher end of the income spectrum with less people in the

middle. As a result, this type of income inequality could affect our understanding of demographics of each district.

The graphs obtained for each district and neighborhood are below. When these graphs are accessed via the html file, they are interactive as you are able to see the district name and/or the neighborhood name of each of the points when you hover over the point. This feature allows the person interacting with the data to gather a bit more insight into the significance of each point.

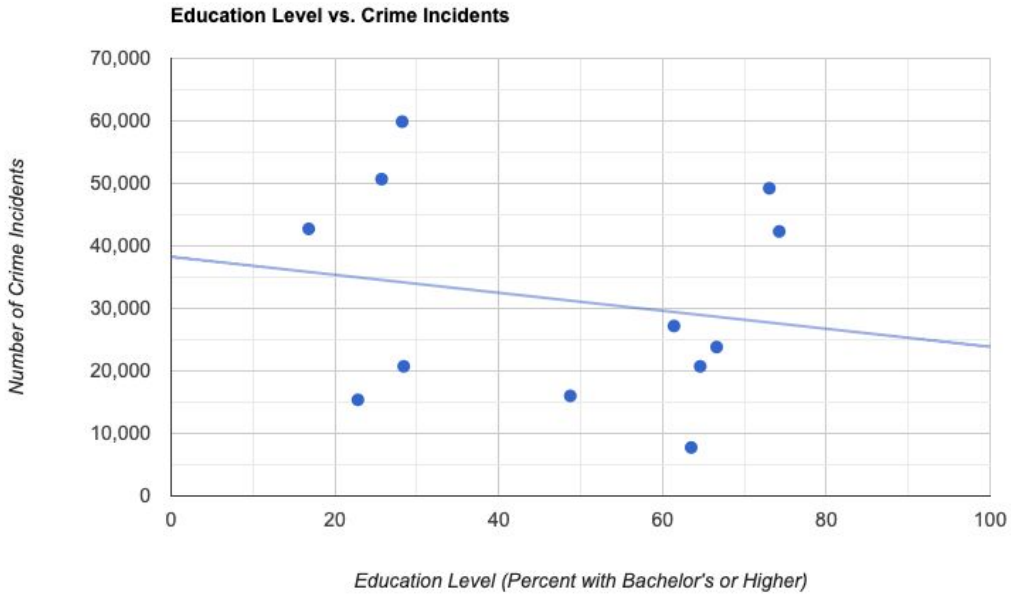

Figure One. Correlation coefficient of 0.1789

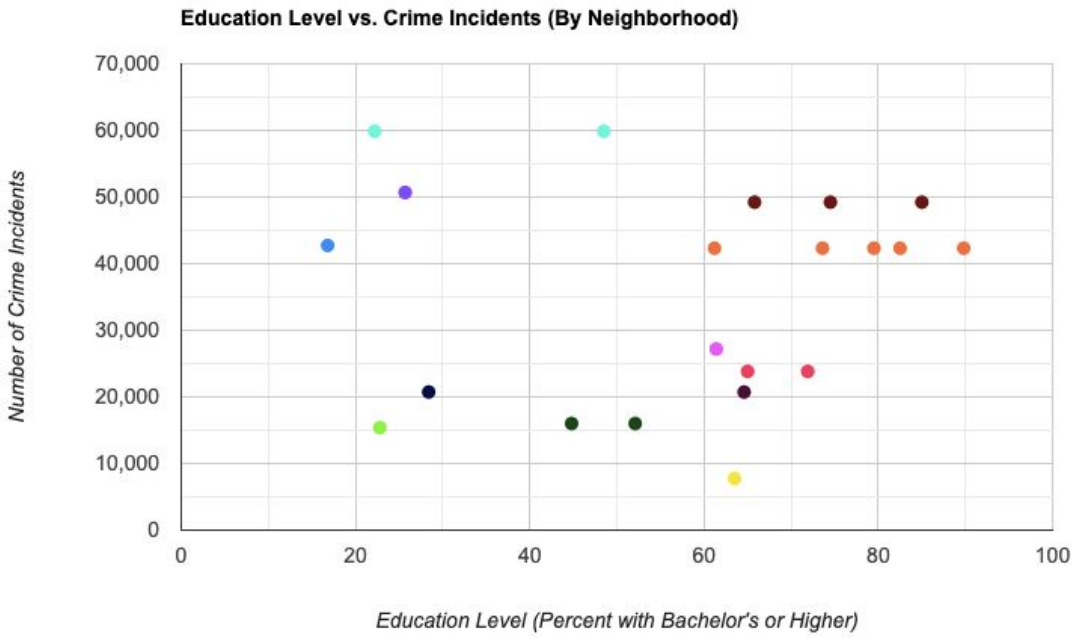

Figure Two.

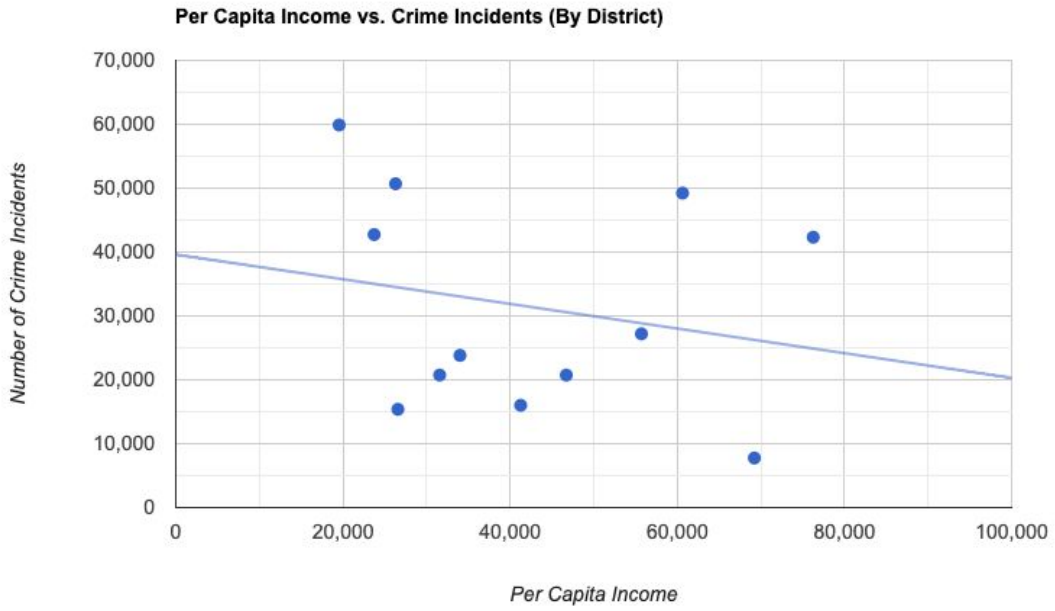

Figure Three. Correlation coefficient of 0.0481

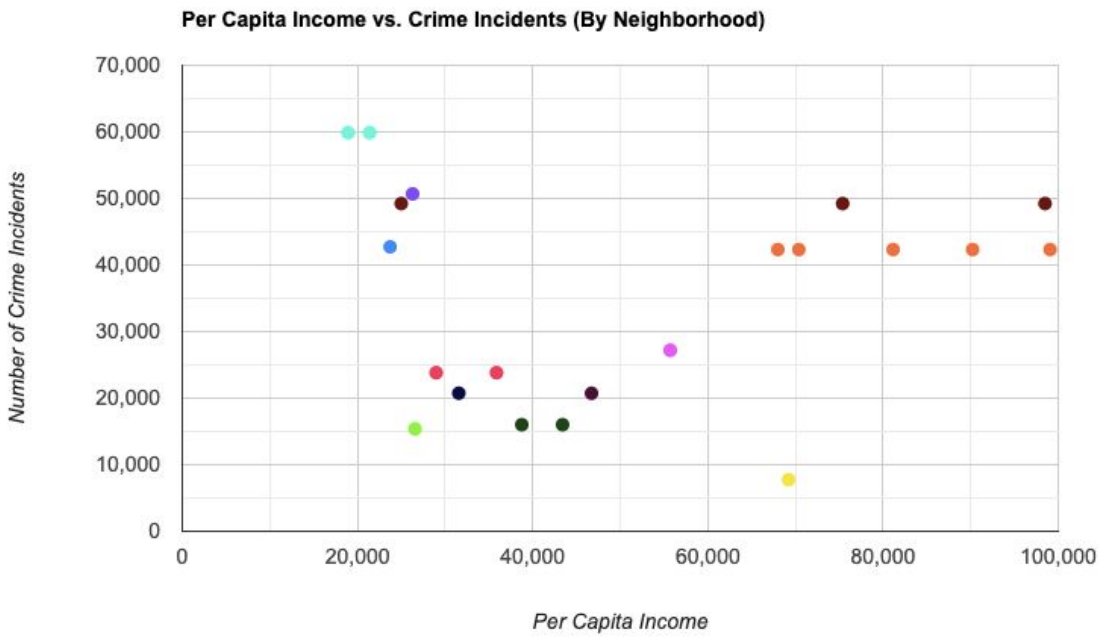

Figure Four.

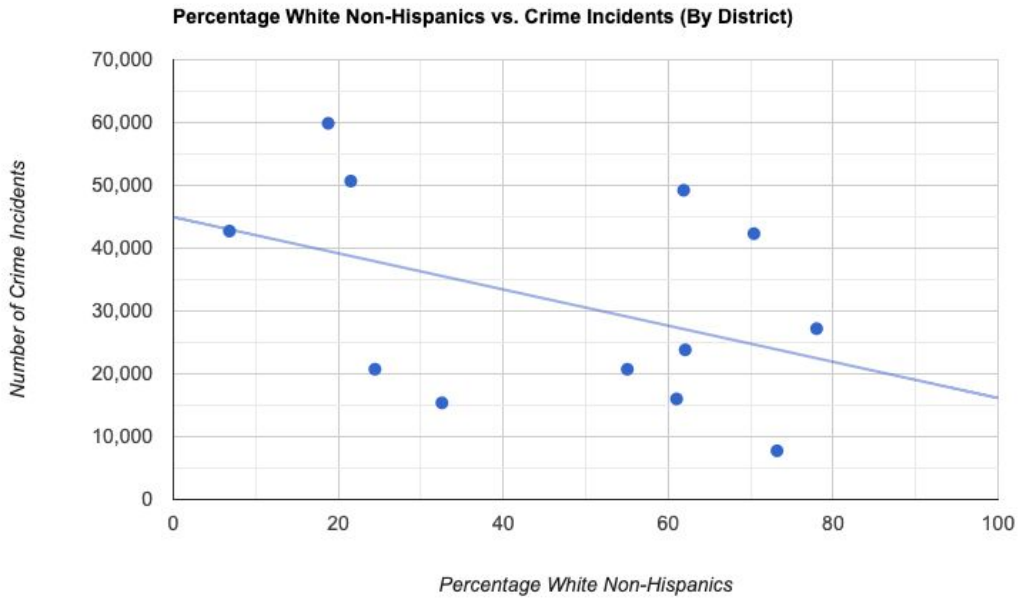

Figure Five. Correlation coefficient of 0.0354

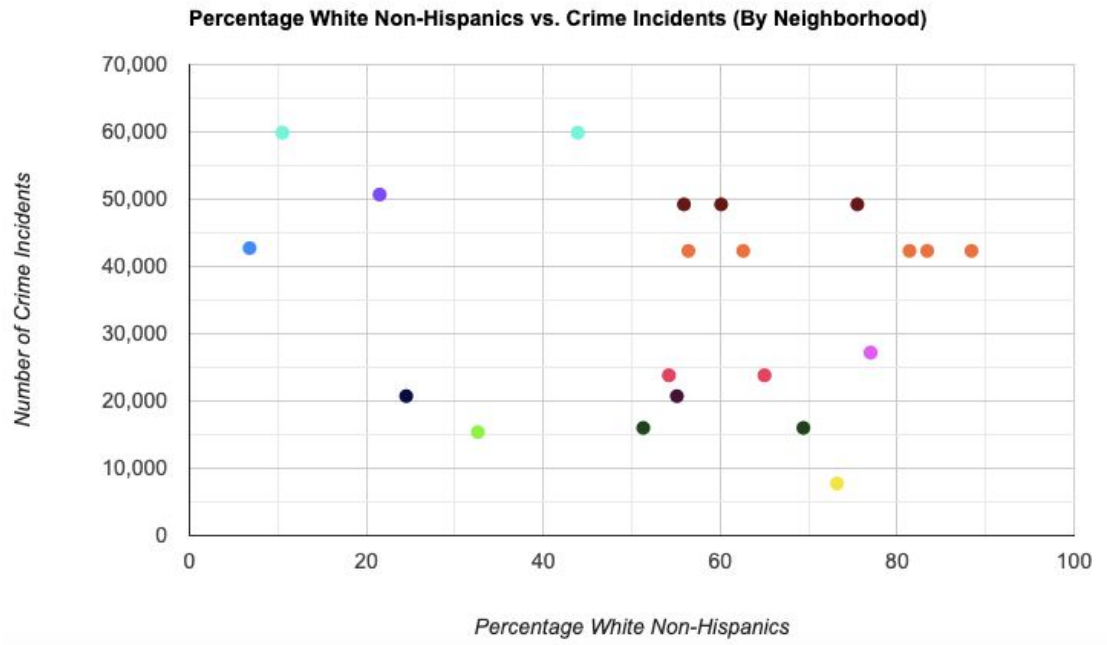

Figure Six.

The clusters produced by K-means were similarly uneventful. For the most part, they are plotted relatively nearby the police stations. Figure seven below is a screenshot of the interactive map displaying crime clusters (as read location markers) and police stations (as blue houses) for each district in Boston. The graph displays information about the district each point belongs to when one is selected and also shows general geographic information about Boston. The map was created by using the procedurally produced .kml files containing the cluster coordinates with Google's MyMaps toolset.

Figure Seven.

Conclusions and Future Work

As it turns out, police districts in Boston are themselves more homogenous than we anticipated, and so we were not able to uncover any critical correlations or insights based on crime centers. However, there are many factors that likely affect how crime centers in a city like Boston might emerge that were beyond the scope of this project. For example, while we didn't find a correlation at the police district level between average per-capita income and crime rates, there could be other ways money affects the way crime is distributed like through business presence or investments in certain neighborhoods. Similarly, it's possible that crime is clustered so near police stations because government buildings are often bought cheaply, and so might be built in

undesirable and potentially more dangerous locations. These are all things that would make good questions for future work.

On that note, there are other loose ends that did not end up making it into our project. We had originally intended to look at the incomes of officers per district to see if there was a connection between the amount of arrests and the amount of money an officer made. We did not get to use the income data, but we cleaned it up so that it would be easy to look at the income data so that we see if there was another factor that affected the number of crime incidents per district.

Another element of the crime incident data that we could have examined is what type of crime was the incident. Vandalism is certainly a much different crime than murder and it could be interesting to reflect that in the data somehow. Perhaps there could be some way that we could “rank” each category of incident so that we could use that new data in calculating where the police stations should be for each district.