Relationship Between Crimes and Service Requests

Arjun Lamba

Boston University CS591 L1 Data Mechanics, Fall 2016

1. Introduction and Data:

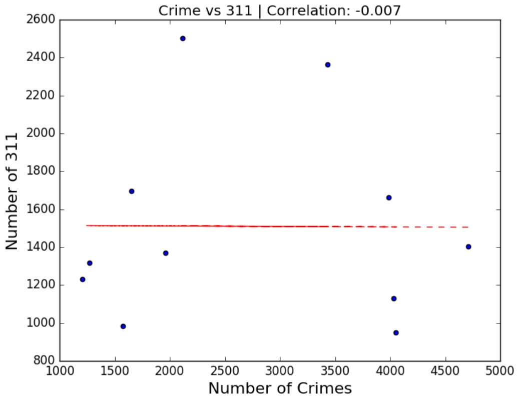
The goal of this project was to look at how the cleanliness of an area is related to the number of crimes that are committed in that area. The motivation comes from the "Broken Windows Theory" which says crime emanates from disorder and if disorder is eliminated then serious crimes would not occur. There are two types of disorders: Physical and Social. Physical disorder is characterized by vacant buildings, broken windows, vacant lots filled with trash and the general dirtiness of an area. So I wanted to see if the areas that require the most upkeep (I am assuming that the more run down a place is the more upkeep it requires so more service requests will be generated) also produces the most crime. To determine how well maintained an area is I looked at the number of 311 services requests, the number of pothole repair requests and the number of mayor hotline requests. In addition to the previous three the crime and development data sets were taken from the City of Boston Data Portal. The crime dataset gave the locations and types of the crimes committed between 2015-2016. The rest of the datasets gave the locations of where the service requests were coming from and the type of requests. The development data gave the locations of Boston's neighborhood development project buildings. This analysis can help with Boston's police policy by giving them an idea of which areas to focus on.

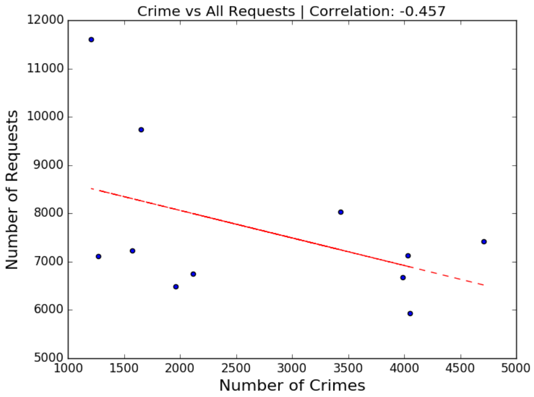
2. Methods/Algorithms:

First I grouped the crimes and requests according to the zip code they occurred in. The problem here was that in the crime data set there was no column for zip codes so I had to map the police districts to the zip codes. There are fewer police districts than zip codes so crimes that should have been in different zip codes were being grouped together. The requests had more zip codes than the crimes data so only zip codes that had crime greater than 0 were considered. Then I computed the Pearson correlation, averages and the p-value on crime vs. 311 requests, crime vs. potholes, crime vs. hotline and crime vs. all (adding up all the requests in a zip code). The second method was to run the k-means algorithm on the locations (longitude and latitude coordinates) of the crime versus the 311 requests. One of the data transformations was to convert the location information into GeoJSON format so getting the coordinates did not require additional processing. Only for the crime data set the first few entries had coordinates (-1, -1) which were thrown out. To get an idea of how the number of crimes and 311 requests were related I looked at the histograms of the locations of crime and 311 requests.

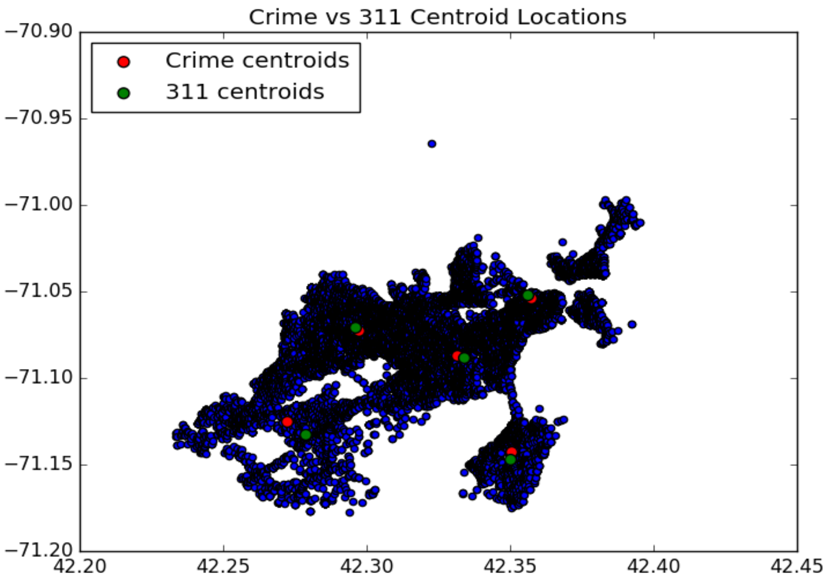
3. Results:

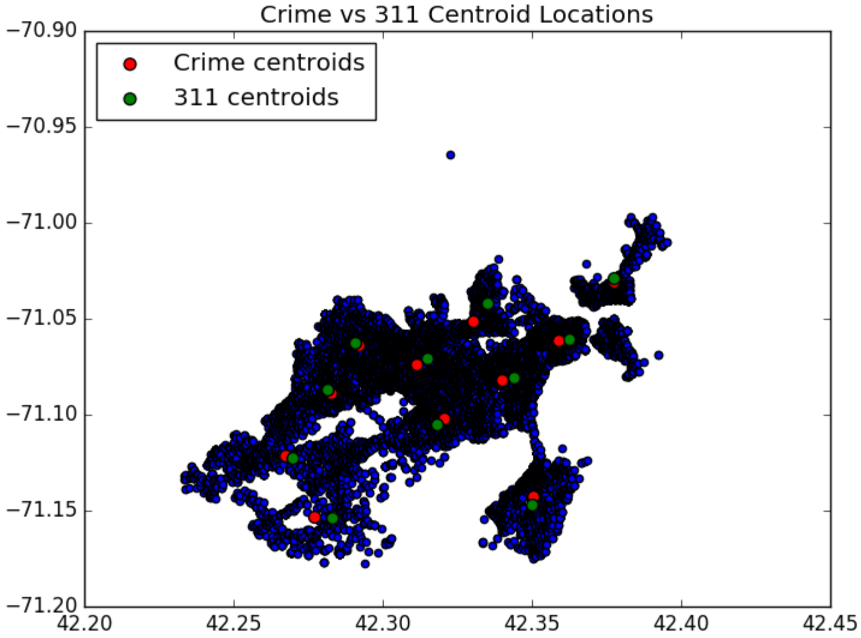
When grouping the crimes and service requests according to zip codes and then calculating the correlation the correlations were all between 0 and -0.5. This means that there is either no relationship between the dirtiness of an area and the amount of crime, or whenever a relationship does exist it tends to be inverse (an area with more requests will have less crime). As mentioned before the crime dataset didn't have zip codes so crimes were assigned to zip codes manually using police district information. This reduces the accuracy of the analysis because crimes might be grouped into the wrong zip codes/neighborhoods which might explain the unexpected correlation results.



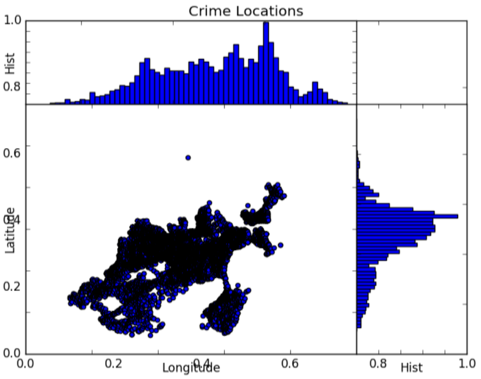


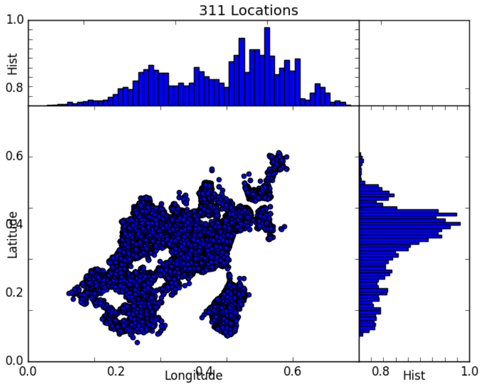
The k-means analysis was done using a range of clusters from 2 to 15. The centroid locations of crime and 311 requests that were generated by the k-means algorithm were in very similar locations. To see if the centroids were related I produced the centroids for 5 and 11 clusters (shown below). The fact that the relative positions of the centroids of crime and 311 didn't change would mean that crime and 311 requests are positively related (area with more service requests have more crime) which was the expected result.





The histograms were first produced to see if there was any common patterns in the distributions of the location coordinates between crime and 311 requests. I wasn't able to do much analysis on them but from the figures the local and global maximum of the longitude and latitude histograms of cime and 311 look very similar. This means that there is a positive linear relationship and support the results of the k-means analysis.





4. Future Work:

These two analysis show that how we aggregate things will have a major impact on how the datasets are related. Cambridge open data crime and service request data sets don’t contain coordinate information (they just list the neighborhood) but according to the analysis above it would be more beneficial to have longitude and latitude coordinates of each entry. In the future I would like to calculate the gap-statistics for the k-means and quantify the centroid locations. I would like to group crime more accurately by using neighborhood boundary coordinates. It would also be interesting to look at the frequency of the types of crimes and their relationship to service requests.