Streetlights and Crime: Characterizing T Stop Safety at Night

Adam Elass
May 4, 2016

Introduction:

The objective of this project was to analyze the safety of T stops at night for a few given lines and, in doing so, to see if a correlation exists between the number of streetlights and crimes within a given radius of a station.

To achieve the goals of this project, the necessary datasets needed, all provided by the City of Boston, were the Crime Incident Reports from the City of Boston (2012-2015), Streetlight Locations for the City of Boston, and MBTA T Stops data.

From these three original datasets, I derived some other datasets, mostly to ease the analysis portion of my project. These new ones were as follows: Stations, which coupled each crime with the closest T stop to it, provided that it is at most one mile away; Crimes and Stations, which went through Stations and kept only those tuples where the crime was within a dist_to_station distance from the station (where dist_to_station is defined in the code, and was 500 ft. for my analysis); Stations and Streetlights, which matched each station with all the streetlights within a dist_to_station + dist_to_streetlight distance from it (dist_to_streetlight was also defined in the code, and was 50 ft. for my analysis); Crimes and Streetlights, which matched each crime with all the streetlights within a dist_to_streetlight distance from it; Crimes and Streetlights per Station, which matched each station with the number of streetlights and number of crimes around it; and Stations and Scores, which matched each stations with its assigned score according to my scoring metric.

Assumptions and Limitations:

The streetlight data I used gave me streetlight locations but not the times the streetlights were on, so I assumed they would turn on at sunset and turn on at sunrise.

Also, regarding the location information provided by the crime data, the documentation states that the x and y coordinates are "obscured to the street segment centroid for privacy", so this may affect the results slightly (e.g., perhaps when I was filtering, a crime was discarded because it's obscured location doesn't fall within the desired distance, even though the actual crime itself may have).

Methods and Results:

As a general overview, my approach was as follows: I initially filtered out crimes that did not happen between sunset and 1AM, since I assumed the streetlights would not be turned on before sunset and most T stations seem to close around 1AM. Next, I selected all crimes that happened with a given distance from a station, and then selected all streetlights that were within a given distance from a certain crime. Much of this filtering and selecting was done by iterating through the elements to see if they should be selected or discarded.

After getting all the data I needed, I was able to count how many streetlights and crimes were around a given station (as mentioned in the previous section, this information was stored in the Crimes and Streetlights per Station collection). Using this information, I calculated the Pearson correlation coefficient to determine whether a correlation between the two existed, and if so, whether it was positive or negative. At the beginning of the project, I assumed that there would be a negative correlation as one would imagine that less lighting would be conducive to

more crime, and conversely, more lighting would mean less crime. However, the calculated Pearson correlation coefficient turned out to be 0.406289747918, with a corresponding p-value of 0.000361666922126. While we can't really comment on the extent of the correlation, we can at least conclude that a positive correlation exists between the number of streetlights and number of crimes (meaning, at least according to my analysis and results, that more streetlights means more crimes). Although this may seem counterintuitive at first glance, it does make sense if we realize that there are more important factors that decide crime rates, such as population size, and these factors could also influence the number of streetlights.

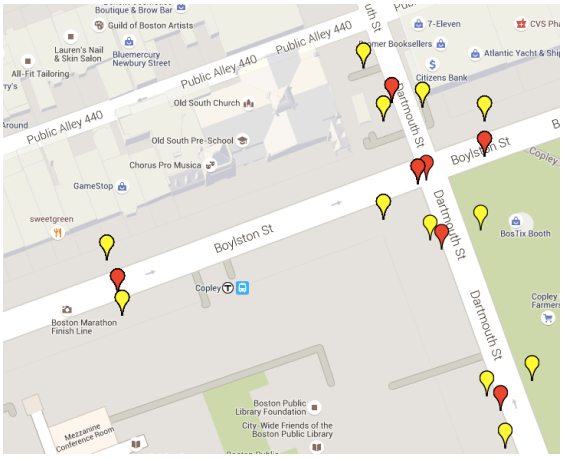
As for characterizing the safety of each stop, I chose to do a weighted sum of the crimes as my scoring metric. To do this, I first went through each station to get the number of crimes that happened at each hour between 9PM up to 1AM (I chose 9PM as the starting hour since different times of the year have different sunset times, whereas throughout the entire year the sun will have set before 9PM). I then assigned weights for each hour, with earlier hours getting higher weights (e.g., 9PM to 9:59PM had a weight of 4, 10PM to 10:59PM had a weight of 3, etc…). My reasoning behind this was that as the hour gets later, the stations tend to get less crowded and hence crimes may be more likely to occur, so a crime that happens at an earlier hour, i.e., one where there would be more people/activity, should be given more weight. I then multiplied each weight by the number of crimes for that hour, and got the total sum which was the score.

For example, if a station had 3 crimes between 9PM and 9:59PM, 5 crimes between 10PM and 10:59PM, 4 crimes between 11PM and 11:59PM, and 7 crimes between 12AM and 12:59AM, then its score would be: 4*3 + 3*5 + 2*4 + 1*7 = 42.

After calculating the scores for each station, the results showed that the station with the greatest sum (i.e., the 'least safe' station, according to this metric) was North Station, with a score of 765. The station with the least sum was Boston University East, with a score of 5.

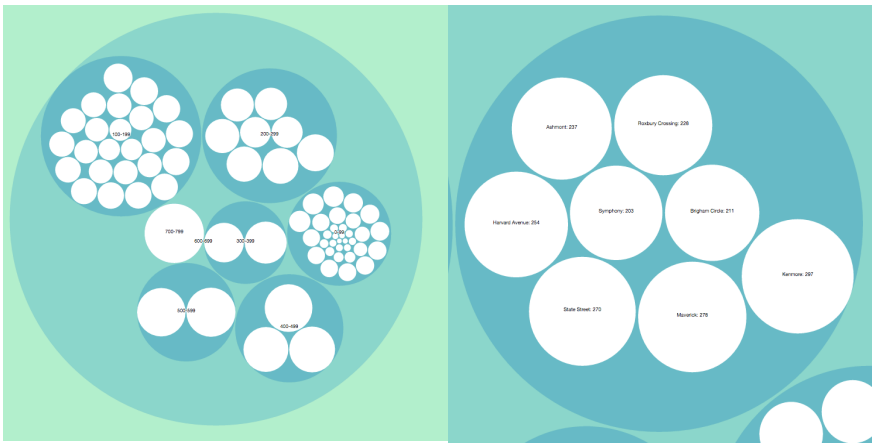
Visualizations:

My first visualization was an interactive map (you can move around, zoom in/out, etc…) of Boston with the crimes shown in red and the streetlights shown in yellow. This was done using the gmplot package for Python 2. An example screenshot is shown below.



My second visualization was a zoomable circle packing visualization for the scores. Stations are grouped according to score ranges, such as 0-99, 100-199, etc… Example

screenshots are shown below. The screenshot on the left is the entire visualization and the screenshot on the right is one of the bubbles zoomed in.



Future Work:

There are numerous things that could be done to improve upon or extend this project. For starters, the crime dataset includes a lot more specific information for crimes, such as the number of perpetrators, the day of the week, description of the incident, etc... Therefore, someone could, for instance, focus only on specific types of crimes and score stations based on that. Or perhaps only crimes that occur during the weekends could be the focus.

Also, if someone wanted to improve the accuracy of the project, some more research could be done on whether or not all the relevant streetlights around a station/crime were functioning at the time of the crime, or even whether or not a certain station was closed for any reason at a given date. Furthermore, other factors can be taken into consideration when characterizing T stop safety, such as number of people in and out of a station during a given time frame.

All in all, there are countless factors that could be incorporated into the analysis and various choices that could be made to perhaps create more detailed or accurate results.

Conclusion:

Although my approach and analysis/results may not have been perfect, I hope my methodology proves somewhat useful or helpful to anyone attempting to perform a similar task. The scoring metric, although somewhat simplistic, did allow for a diverse characterization of the different T stops, and the Pearson correlation coefficient calculation did allow for a conclusion to be reached regarding the relationship between the number of streetlights and number of crimes.

In the end, I was able to answer the question of whether or not streetlight and crime numbers are correlated, at least based on my data. Also I succeeded in achieving the objective I set out to achieve, which was to analyze and characterize the safety of T stops at night. Therefore, I believe this project and its results were successful in that regard.

Dataset sources:

- Crimes for the City of Boston (2012-2015) API endpoint: https://data.cityofboston.gov/resource/7cdf-6fgx.json

- Streetlight data: https://data.cityofboston.gov/api/views/fbdp-b7et/rows.geojson

- MBTA Stopsbyroute API endpoint: http://realtime.mbta.com/developer/api/v2/stopsbyroute

Source for code for zoomable circle packing visualization:

- http://bl.ocks.org/mbostock/7607535