Streetlight Maintenance Priority Queue

Matthew Endo

Technical Method

Data Processing

The datasets that were used were the “Open Get it Done Requests” and the “Get it Done Request closed in 2016-2022” (accessed on 9/22/2022) from https://data.sandiego.gov/datasets/get-it-done-311/. Additionally, US 2010 Census data was used via uszipcode python package. All python code can be found https://github.com/matthewendo7/SD_streetlight_maintenance_priority_queue.

All seven datasets containing closed requests were combined into a single dataset. The closed requests dataset was filtered of all requests that did not pertain to “Street Light Maintenance”. Similarly, the open request dataset was filtered of all non-“Street Light Maintenance” requests. The closed requests dataset contained 45603 entries and the open request dataset contained 7949 entries. Of the 45603 closed requests entries, 791 entries were referred to other associations (Caltrans, Parks & Recreation, utility companies, etc.). The referred entries were removed due to lack of information regarding the completion of the requests. The closed requests dataset had 44812 entries remaining.

Of the 44812 closed requests entries, 7980 entries were either missing a zip code or contained a zip code outside the San Diego County area. Furthermore, of those 7980 entries, 51 entries did not have coordinate data to determine location. Those 51 entries were removed due to difficulty to systematically determine location. Of the 7949 open requests entries, 19 entries were either missing a zip code or contained a zip code outside the San Diego County area. All 19 entries did contain coordinate data to determine location. The uszipcode Python package was used to determine zip codes from coordinate data so that all entries in the closed requests and open requests datasets now contained zip codes.

Priority Determination and Priority Queue

With the readily accessible data, three factors were chosen to determine priority for requests: when the request was placed, crime rate in the location of the streetlight, and foot traffic in the location of the streetlight. Actual crime rates for each San Diego neighborhood were difficult to access. In lieu of crime rates, median household incomes from the 2010 US Census via the uszipcode Python package were used to represent crime rates for the various zip codes. Free and accessible foot traffic data was difficult to access. In lieu of foot traffic, population density was used to represent foot traffic data. The uszipcode Python package was utilized to determine population density from the 2010 US Census for the various zip codes.

To determine priority, a priority score (from 0 to 1) was determined for each of three factors for each entry. The total priority score was determined by 40% from the priority score for when the request was placed, 40% from the priority score for crime rate (as represented by median household income), and 20% from the priority score for foot traffic data (as represented by population density). The factor for when the request was placed was normalized where the oldest request (2318 days from today) received a 0 and newest request (3 days from today) received a 1. Similarly, the factor for crime rate was normalized where the zip code with the lowest median income (and likely higher crime rates) received a 0 and the zip code with the highest median incomes received a 1. Finally, the factor for foot traffic was normalized where the zip code with higher population density (and likely higher foot traffic) received a 0 and the zip code with the lowest population density received a 1.

The priority queue was constructed from the heapq Python package because constructing the priority queue from a heap data structure allows fast enqueue (O(log(n)) and dequeue (O(log(n)). The total priority score and the service request id for every entry was inputted into the priority queue. Entries are dequeued from lowest total priority score.

Potential Quality Issues

Data Quality Issues

Potential additional factors include number of malfunctioning streetlights per report and whether the streetlight is completely out, blinking, on during the day, etc. Some of this information can be pulled from the "public_description" column in the dataset, but it is not easily systematically acquirable. Missing or incorrect zip code data required using the uszipcode package to determine zip codes for reports missing them.

Assumptions

The first major assumption is that the user-inputted data (specifically latitude, longitude, and zip code) are true. A few sanity checks were performed such as latitude being between 32 and 33.5, longitude being between -118 and -116, and zip codes being between 91900 and 92200, but nothing extensive.

The next major assumption is that the uszipcode Python package is providing the correct zip codes based on the coordinate data. Similarly, that the package is also providing the correct 2010 US census data corresponding to the zip codes.

The final major assumptions are that higher crime rates do correlate with lower median household incomes and higher foot traffic correlates with higher population density in the San Diego area. There a multiple publications that report this correlation for the world (https://journalofeconomicstructures.springeropen.com/articles/10.1186/s40008-020-00220-6) and the US (https://bjs.ojp.gov/content/pub/pdf/hpnvv0812.pdf). However, it is still an assumption as this priority queue utilizes lower median household income to represent higher crime rates. The higher foot traffic correlating with population density is a more substantial assumption as there was little published evidence to support the assumption.

Results

Key Calculations

Priority scores for each factor (age of report, crime rate, or foot traffic) were calculated by normalizing based on the max and min within the factor:

$$Priority\ Score_{factor} = \frac{factor\ value_{report\ entry} - \ factor\ value_{min}}{factor\ value_{max} - factor\ value_{min}}$$

Total Priority Scores were computed such the score was comprised of 40% from the Priority Score for case age, 40% from the Priority Score for crime rate, and 20% from the Priority Score for foot traffic.

$$Total\ Priority\ Score = Priority\ Score_{case\ age} + Priority\ Score_{crime\ rate} + 0.5 * \ Priority\ Score_{foot\ traffic}$$

Simulation of the priority queue was performed and compared to the statistics from the actual closed reports. To create the simulation, the actual closed dates were assumed to be the same closed dates for the simulation. All the streetlight reports (open and closed) were put into a single dataset and the priority queue was developed for the entire dataset. The closed dates were sorted and assigned to reports according to the priority queue (i.e. the first report in the priority queue was assigned to the earliest closed date). With that the simulated case age was calculated based on the assigned close date by subtracting the request date.

The actual closed reports had an average case age of 152 days (89 days median). For the simulation, the average case age was 133 days (33 days median). By placing some focus on completing older reports first, this helps to address equity. To gauge the potential impact on public safety, the average and median crime rate priority score for the actual closed reports and simulated closed reports. The actual closed reports had an average crime rate score of 0.204 (0.167 median) while the simulation had an average crime rate score of 0.173 (0.150 median) where lower scores correspond to higher crime rate. A similar calculation was performed with foot traffic score. The actual closed reports had an average foot traffic score of 0.494 (0.520 median) while the simulation had an average score of 0.443 (0.373 median) where lower score corresponds with higher foot traffic. Emphasizing foot traffic would help to increase would help to address greater public impact as more individuals would be affected with fixing streetlights in higher foot traffic areas.