Street Light Repairs: Performance and Prioritization

Process and Methodology Documentation
Brent Goode

Introduction

The Transportation Department has approached us for help improving their process responding to requests for streetlight repair. The department currently has no systematic way of addressing reports. They would like to emphasize both equity and public safety in a new, systematic approach.

First we discuss issues with the data sources, the steps taken to clean them, and reasons for not including other data at this time. Next we will examine the past performance on repairing street lights to determine how well the department has been doing and to see if there are any celar favoring of some areas over others to ensure equity. We will also examine the department's performance in underserved areas as another way of getting at equity.

Then we will examine the correlation between street light problems and crime and other safety related data to get at the safety aspects of the problem. Finally based on what we find here we recommend a prioritization methodology for addressing street light repair requests. We also recommend improvements to data collection and data sets that would allow improved performance measurement and prioritization methodologies.

Data Sets: Cleaning, Shortcomings and Issues

The primary data set used in this analysis is the Get it Done reports dataset available on the city's website. These reports are already well formatted, internally consistent, and largely filled in so little cleaning was needed. When reading and combining data from multiple files of closed records the zip code field ended up as a mix of text and float. This field had to be converted to a consistent type across all data. Additionally, the field with date and time the request was made had to be converted from text to a datetime type for some calculations. Many requests for street light repairs were not closed by the Transportation Department, but were referred to other agencies. When filtering Get it Done reports for street light repairs, reports that were referred were also filtered out.

Each report has a free text field called public_description where the submitter describes the report they are submitting. The contents of these fields lead to some concerns about the completeness and accuracy of this data. The reports are tagged with a timestamp when they were submitted, and we assume that this is shortly after the issue began, but a quick look at the free text fields showed one with the following quote, “Light has been out for a very long time.”

The length of time this report has been waiting for service is an underestimate of the time that the light has been broken, but how much of an underestimate cannot be determined, as “a very long time” cannot be quantified.

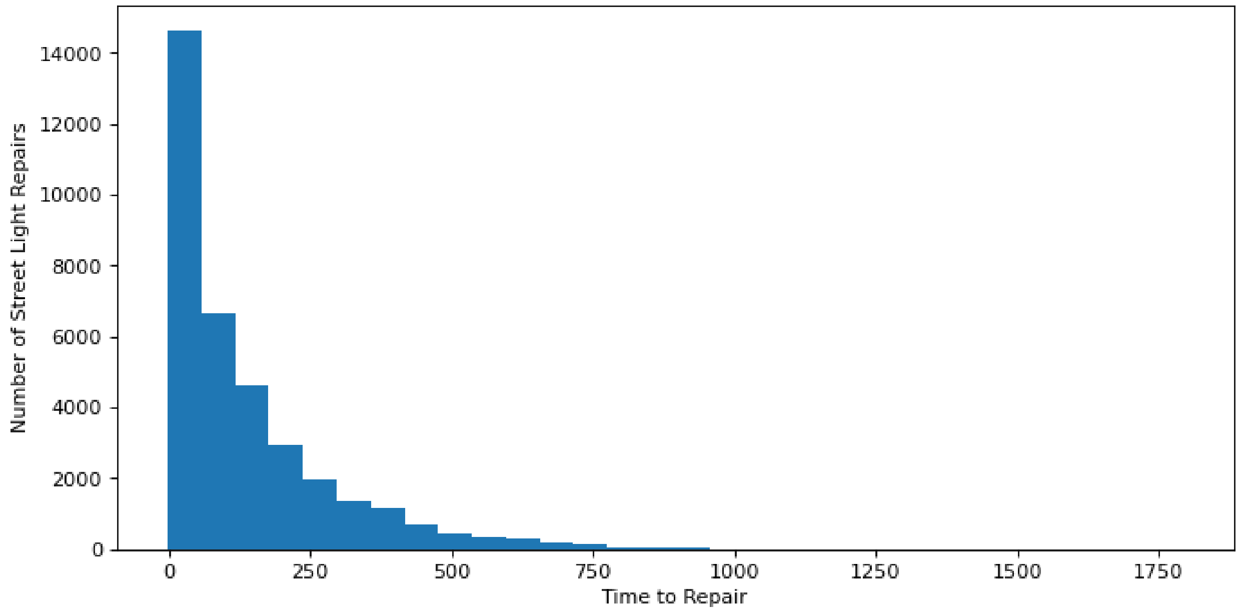
The Get if Done data on closed reports is organized into files by the year when the report was closed. The first request in the dataset is dated 5-20-2016, so the 2016 closed file is not a complete year of data. In addition it is impossible to close a request that has been open for longer than the request system has been in operation. The very earliest records have closure times that are necessarily shorter than the full span of real closure times. They were fixing lights during this year that had broken before the system started recording data, but these repairs are missing from the files. For this reason we only consider reports of repairs that were closed in 2017 or later.

When submitting reports in the app the phone’s location is taken as the starting location of the problem. This would result in accurate location information if the users did not adjust the pin’s position, but not all reports originate from phones. Some are reported by repair crews and these can also be assumed to have accurate location information, but none of the open street light reports have this origin. There are categories for reports submitted via web, email, letter, and walk-in. The location on these report types cannot be trusted as much as locations auto generated by the mobile app.

Email reports of current street light issues are less than a tenth of a percent of the total and phone reports are about five percent of the total, so location issues here are acceptable. However, web reports are thirty six percent of all open street light requests. If these have significant problems with their location it could impact our results. Inaccuracies in location would not only affect correlation with other reports, but also the assignment of reports to geographic regions such as council districts and zip codes.

Related to issues with location is the assumption that multiple reports about the same street light are correlated correctly and all are closed when the light is fixed. If not then some of the long outstanding requests could be related to lights that have already been fixed. From the outside it is not clear how reports are correlated against others to determine which are duplicates. Therefore we must assume that multiple reports of the same issue are correctly correlated.

With this all said, we can do some preliminary examinations of the distribution of current request ages and historic repair times. The figures below show the historic distribution of repair times for repairs closed from 2017-2021.



Because the distributions of repair times and current request ages are not close to normally distributed, median is a better single metric instead of mean.

There is no baseline data set of all street lights, working or broken, to baseline the numbers broken against. A very useful metric for determining equity would be the fraction of lights in a geographic area that were out at any given time. But since we cannot calculate this we must rely on median time to repair and age of open requests as our measures of past and current performance, respectively.

In order to address equity we need additional information on which parts of the city are disadvantaged or poorer. One good definition of the most historically disadvantaged areas are the Promise Zone. “Promise Zone is a federal designation for an area of San Diego that has high unemployment, poverty, low educational attainment, insufficient access to healthcare and healthy foods, rising crime rates, and the least affordable housing in the country.” The geographic definition and source data on this zone can be found on the city’s website.

The data on wealth and poverty used in this study are the median household income by city council districts for 2020 is published by the San Diego Association of Governments (SANDAG). The report for the first city council district is here. Other districts can be found by changing the number at the end of that link between 1 through 9. Since those reports were in PDF format, the data on median income was manually collected and written into the analysis code. These reports also provide demographic information on the districts, but examination of these factors was skipped due to time constraints. More fine grained data on income and demographics by zip code and community would also be useful for a longer study.

Finally to address safety, crime report data would be most useful. Specifically reports of crimes that happened at night near the location of a broken street light after that street light was reported as not being on. The first crime report database discovered was the crime report data that SANDAG publishes on their website. The location data in this report is a text field that only records the block or intersection. In addition there is a high degree of variability in how this is recorded. For example there are several ways of recording boulevard: BLVD, BL, or BOULEVARD. The Get It Done dataset does have a text street address field, but these are far less clean than the latitude and longitude data.

A first attempt at cleaning out the block part of the crime report text was made to link the two datasets based on text matches, but this does not account for location that is recorded as being at an intersection. In the end the need to complete this work on a deadline made using this data set impractical. Going forward if crime data is needed coordination with local law enforcement to get access to data that they do not publicly publish with better location data is a more productive strategy.

Determining the statistical significance of the difference between repair data for different categories that will be used in this report is not a trivial problem. First the distributions of repair times are clearly not normal, which could be dealt with. The larger difficulty is that we cannot say that any one observed repair time is independent from the rest, which is a usual requirement of statistical tests for significance. If a crew goes to repair a street light, it is very likely that they will fix more than one on that day and will focus on other lights close to the first. This would place all the repairs done that day in the same geographical area even if the later repairs were on light that had been broken for a short amount of time.

These challenges to determining statistical significance are not insurmountable, but the limited time available in this challenge (and in a normal work day) would be better spent examining other topics that would produce quicker results. Especially since in presenting results like this to leadership (city councilors for example) they are unlikely to be satisfied with the explanation that their district's poor performance should be ignored because it wasn't poor enough to cross some statistical threshold.

Current Backlog and Past Performance

Examining the current backlog started with producing a histogram of the age of currently open requests. The median backlog time was calculated for various divisions of the data, council district, zip code, community plan name or community for short as well. These served as the basis for conclusions about the current backlog.

Presenting partial year data on 2022 closure is also difficult. There should be some allowance made for the fact that the measures are subject to change as the year continues to progress. Because of this and the issues with the 2016 closure data discussed above the historic data used was all closures from 2017-2021. The same median request lengths used for the current backlog divided into the same categories (community, council district, etc.) were calculated. The change in the median time to close requests was calculated for each year for the entire city and the various geographic divisions. The median over the entire period for each division was also found.

We considered a calculation of the number of open requests on any given day over the past few years. This would have complimented the plots showing the growth in median time to repair over the same period, but not really explained why either was trending the way it was. To highlight the reasons for the growth a different calculation was done. This plotted the number of requests opened versus those closed each year to show that the number of requests closed each year did not keep up with the number of requests coming in. This plot also nicely illustrates that the number of requests coming in had not changed substantially over this period, so an increase in demand can be eliminated as a cause.

Equity Examination

First a definition of equity was needed, and the one given on the Department of Performance and Analytics site was the obvious choice. The search of the Performance and Analytics Departments public data catalog for any useful data turned up the Promise Zone as a good way to define traditionally disadvantaged areas. In addition income would be a measure of inequality that could be examined in relation to longer repair times. Finally, there is the self-referential inequity in the rate that street lights have been repaired in the past for neighborhoods across the city. This inequity may or may not be tied to poverty or other disadvantages, but if it is corrected it can result in a more complete leveling of service than any other factor.

The first examination of equity involved separating all data into the requests located in the Promise Zone and those in the rest of the city. The functionality of the GeoPandas python library was the key to this separation. Instructions and code examples from the tutorial located here were used to develop python code to perform this separation. The code written to do this was packaged into a function for use multiple times in the analysis code and reuse in later analysis. That function can be found in the project repository here. The result of looking at past repair times and current backlogs showed that the Promise Zone gets slightly quicker street light repairs.

Next the differences in repair performance between different city council districts was examined with an eye toward the different median household incomes of those districts. For this and many other examinations by category the python Pandas “grouby” aggregation operator was used to find the median repair times and counts of repairs for points in different council districts, zip codes, and communities. Data on which of these geographic areas repair requests were in was already provided by the request themselves. The results by council district was joined on the hand entered data on median household income and scatter plots were made. The past repair times had some bias toward slower repairs in poorer council districts, but the current backlog has the shortest age of open tickets in the poorest two districts.

One issue with this analysis is that the median income by council district hides some large differences within districts. Zip code and community are more fine grained geographic areas that may provide better resolution on this issue. One difficulty in doing this analysis is that some of these divisions had very few closed requests over 2017-2021 or in the current backlog. The median closure time for these small areas included some outliers, so a threshold of 10 requests was set as the minimum to include a zip code or community in this analysis.

Safety Factor

Clearly there are some street light repair requests that have an impact on the safety of the area they are in. A quick survey of the user comments that are submitted with request found the following quote, “we need to get our street light fixed, we have no light out here during the night and due to that he have been robbed multiple times please do something about it!!” If the crime reports on these robberies could be linked in space and time to the street light repair it would provide powerful quantitative evidence of the safety impact.

Put more generally, the number of nighttime crime incidents in the vicinity of a light after it was reported broken would be a useful measure of how that light’s malfunction had impacted safety. However the crime data that is publicly published does not have easily usable location information. As a substitute other Get it Done reports data can be used to demonstrate what an analysis like this might produce. In addition to street light issues there are other categories of reports on issues that might make people feel unsafe to be in a neighborhood. These include graffiti, illegal dumping and homeless encampments. These issues plus broken street lights would cause a reduction in foot traffic which would make the area less safe becoming somewhat of a self fulfilling prophecy.

The distance between reports was found using the GeoPandas library. This library can take a detailed geodetic system as an input in order to find very accurate location information. We used this library with data given in latitude and longitude with no specification of the geodetic system. Since no coordinate reference system was given the distance returned was in unspecified units. Working with another program to compare the results returned to the distance between select points it was found that the return was an arc angle in degrees along a cylindrical projection at a latitude near the points. The simplest correction factor to get from this measure to a distance would be:

$$R_{earth} \frac{\pi}{180} cos(latitude)$$

with the units of distance decided by the units for the radius of the earth.

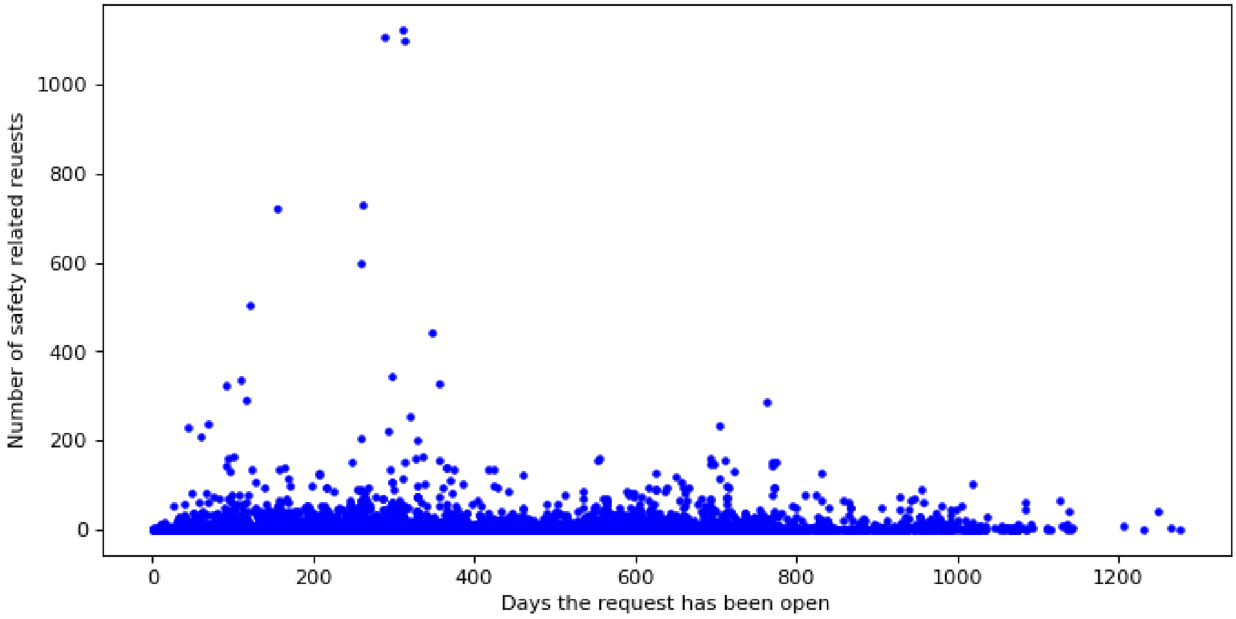
A single fixed value for this correction factor was found and used in the simple demonstration code. This was 307,000 ft per degree. Technically the factor is slightly different for the north end of the city compared to the south. Given the concerns already discussed above about the location information of these reports and the uses to which they are being put, the very fine accuracy that a more complex model would allow are not necessary.

Then the data on open street light requests was looped over to first select safety requests that had been submitted after the individual street light request. Then in a parallelized way using the Pandas package, the distances between the single street light request and the safety related request was used to filter and count the safety related requests. The count was then appended to the data on the street light request.

Looping over the data on street lights was very inefficient compared to a vectorized method of finding time correlation and appending results. But the first pass at doing this would have resulted in not vectorizing the second calculation of distance. It may be possible to do both, but since this calculation was only to be done once for this demonstration, the time spent optimizing this code would exceed the computational time to run it. If this code is ever to be run again, the optimization needs to be done.

One final note is that some street light requests did not concern the light being out. For example some are for lights that are on all the time. Requests like these are not a safety concern so any street light request that did not have a 'service_name_detail' of 'STREET LIGHT OUT' or 'POLE KNOCK OVER/DAMAGE' had its safety factor set to zero.

One concern with this methodology is that the number of safety requests would just become a proxy for the age of the street light request. The longer a request is open the more time there is to have other requests submitted in its vicinity. To check if this happened we look at a scatter plot of the safety requests versus the age of the request. That is shown below. Encouragingly, selecting requests based on safety request count would pick out a subset of younger events and almost none of the oldest ones.



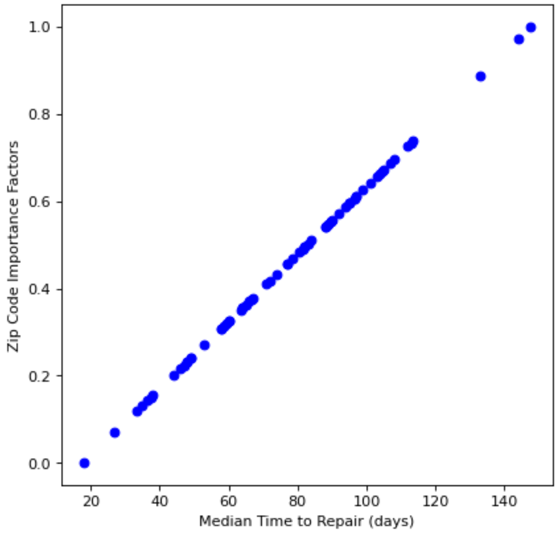
Prioritization Methodology

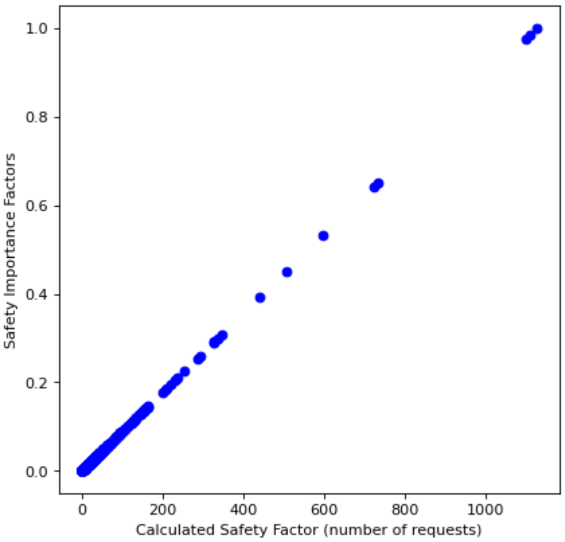
The analysis so far suggests a number of factors that could be used to prioritize street light repairs. Because each factor is quantized with a higher number recommending a higher priority, a weighted mixing of the factors into a single priority number can be calculated to do this. Deciding on the weighting to rate the relative importance of these factors is beyond what analysis is capable of as this ranking involves balancing the desired outcomes against each other. Put another way, the decision to raise the importance of one factor lowers the importance of the others. How much leadership is willing to trade off one factor in order to get more of another is their decision.

But first the various factors need to be converted to the same scale. Age of request and geographic past performance are measured in days, but the range of variation of each factor is not the same. Being in or out of the Promise Zone is a binary factor. The safety factor is a count. The value for each instance of a factor *i* is used to compute the scaled factor for that instance according to the following formula

$$scaled\ factor_i = \frac{factor_i - factor_{min}}{factor_{max} - factor_{min}}$$

The result is a scaled factor that spans the range of zero to one with each individual instance linearly scaled so that they fall proportionally in between. This will allow each factor to be multiplied by a weight and added to the others to get a single priority rating that reflects the full span of each factor's variability. An example chart showing this scaling for two factors, zip code and safety are shown below. Each point on the left is one zip code and on the right is one repair request.





There are special cases of zip code and community divisions that had fewer than 10 closed requests over the time examined. Previously we discussed that the median closure time for these areas were outliers. Those were not used to find the minimum and maximum values in the formula above. Instead a neutral weight of 0.5 was given to any zip code or community that had fewer than 10 closed requests. Also, since being in the promise zone is a higher priority than being out, being in is rated as 1 and being out as 0. Once this is done, a user input priority weight for each factor is multiplied by the scale factors for all the requests and the results are summed into a single number that gives a higher priority for higher values. The current code writes this to a file as a report, but a visualization tool like the Spyder IDE data explore gives a better view of the results.

This scaling and prioritization scheme for geographic areas would eventually produce equality across all areas of the city. For example, by prioritizing repairs for the next year in zip codes that had long repair times in the past, the points on the top right of the graph on the left would be brought toward the center. By being deprioritized those on the bottom left may also move toward the center. If this process was repeated each year with the last year’s data then there would be an iterative process of moving zip code repairs times toward the city wide median, resulting in geographic equity.

Recommendations for Future Work

Work with law enforcement to get crime related data with better geolocation data.

Develop a database of all street light locations if one does not already exist.

A useful study to supplement this work and any others that rely on the Get it Done report dataset would be to examine the geographic distribution of all reports versus the population density of various areas of the city. Systemic differences in the number of requests coming in from different neighborhoods could be driving systemic differences in service levels.

Create a prioritization tool that not only sorts repairs according to the methodology above, but also selects a set of other repairs to go with those based on proximity and their own priority. Determining the number and type of requests to put into this work package would require close coordination with the Transportation Department to learn how they group their work for maximum efficiency. Understanding the working environment and practices of the Transportation Department would help in developing this tool. This could be achieved by having analysts join crew for a time to directly observe the constraints the crews have to work under. Also, knowing that the analysts had taken their views into consideration would make the workers of the Transportation Department more accepting of the tool.