Evaluating the Effectiveness of the 311 Response System
City of Boston 2020-2021 311 Report Data

Executive Summary

When looking at large-scale data, the assumption is that areas with lower median household income (MHI) will experience unfavorable disparities compared to their higher MHI counterparts. This typically presents as fewer public resources or funding available for that population, which can cause a chain reaction leading to more severe consequences. Hypothetically, if an area lacks the resources to repair road defects in a timely manner, drivers and pedestrians are at a greater risk for accidents. If the road defect impacts vehicles that drive over it, car owners will be forced to repair their vehicles more frequently, which will take personal time and financial resources to resolve. Additionally, if the road defect compels drivers to swerve around it (deep potholes), the roads become less safe for any vehicle or person in its vicinity. Higher accident rates can cause higher insurance premiums across the board and anyone directly involved in a collision will have their personal rates raised, and in the event of an injury, will once again have to allocate personal time and financial resources toward resolving any issues. Small-scale events have the capability to lead to broader, more serious issues such as driving up the cost of living to the point of gentrification.

This can be a way for inequities to persist, even if an organization's methods and responses are equitable on paper. A proven example of this is the climate change crisis. While it is true that we all reside on Earth and should be equally affected by the impacts of climate change (such as flooding, storms, and poor air quality), Environmental Protection Agency (EPA) analyses[1] have found that underserved communities tend to face more harm. The study finds that the populations most vulnerable are based on education, income, age, and race/ethnicity. The groups impacted do not have the financial or temporal abundance needed to quickly recover from these large-scale climate events, reducing their overall quality of life; there is less time to care for physical and mental health and less access to the means that make it possible to recover, a concern detailed by WHO[2]. Additional information regarding other potential impacts can be found in the SAMHSA July 2017 Supplemental Research Bulletin[3]. While these communities may be treated equally, the lasting impacts of these issues reveal inequities, and though these inequities may not be caused by climate change or nonemergency issues, it is important that either preventative measures or responses are timely and adequate to minimize any lasting impacts.

---

[1] EPA Report Shows Disproportionate Impacts of Climate Change on Socially Vulnerable Populations in the United States | US EPA," US EPA, September 2, 2021, https://www.epa.gov/newsreleases/epa-report-shows-disproportionate-impacts-climate-change-socially-vulnerable.

2 World Health Organization: WHO, "Urban Health," *Www.Who.Int*, October 29, 2021, https://www.who.int/news-room/fact-sheets/detail/urban-health.

3 "Greater Impact: How Disasters Affect People of Low Socioeconomic Status," SAMHSA - the Substance Abuse Mental Health Services Administration, accessed April 18, 2023, https://www.samhsa.gov/.

Introduction

The initial purpose of this analysis was to uncover and address any inequities found within the Boston neighborhoods. By being proactive and reallocating funds in an equitable manner, it is possible to counter the adverse impacts of inequality within the city. However, as the analysis process furthered, it became evident that most of the Boston neighborhoods shared similarities in regard to their 311 reports. In fact, when observing the attribute “general_category”, it is seen that while neighborhoods have different amounts of opened reports, the proportion of each general category is consistent. The chart shows the distribution of all of the report's general categories by neighborhood by proportion. The n variable is the total number of reports within each neighborhood.

neighborhood	n	Environmental	Infrastructure	Other	Parking	Public Safety
Allston	13652	43.04131	12.08614	12.97978	17.78494	14.10782
Back Bay	26392	42.42952	13.02667	12.02258	18.45635	14.06487
Bay Village	3200	42.68750	12.21875	12.15625	18.06250	14.87500
Beacon Hill	17701	42.35919	12.58121	12.12926	18.54697	14.38337
Brighton	28539	43.23207	12.15179	12.56526	17.53040	14.52048
Charlestown	22729	41.47125	12.52145	12.82943	18.24101	14.93686
Chinatown	4864	42.70148	11.94490	11.67763	19.83964	13.83635
Dorchester	101871	42.24559	12.45399	12.99683	17.90500	14.39860
Downtown	18405	41.14099	13.20837	12.60527	18.58734	14.45803
East Boston	40183	42.09243	12.92089	12.53266	18.19177	14.26225
Fenway	13866	42.88908	12.00779	12.61359	18.28934	14.20020
Hyde Park	25058	42.71291	12.20369	13.04174	17.44752	14.59414
Jamaica Plain	32477	41.33079	12.81830	12.69514	18.11436	15.04141
Leather District	899	42.82536	13.68187	12.12458	15.35039	16.01780
Longwood	1562	40.39693	11.52369	12.16389	20.67862	15.23688
Mattapan	16502	42.38274	12.11368	13.38626	17.50091	14.61641
Mission Hill	8444	44.50497	12.58882	12.16248	17.39697	13.34676
North End	18784	43.43058	12.37756	12.30835	17.57879	14.30473
Roslindale	21813	42.08041	12.48338	13.18021	17.85632	14.39967
Roxbury	44809	42.46468	12.30333	12.60907	17.78437	14.83854
South Boston	57193	42.50345	12.17806	12.48405	18.29944	14.53500
South Boston Waterfront	4029	41.62323	12.75751	12.06255	18.61504	14.94167
South End	48228	42.37580	12.39322	12.79340	18.18238	14.25520
West End	3002	42.79919	13.01072	13.01072	18.42944	12.74993
West Roxbury	19539	41.31737	12.26265	13.31696	18.21485	14.88817

Here you can see that environmental reports were approximately 42% of the report types, infrastructure was roughly 12%, parking was ~18%, and public safety was ~14%. These statistics hold true for every observed Boston neighborhood regardless of variety in the overall number of reports, geographic location, and demographic data.

The variable data remained consistent throughout all neighborhoods, and so, the focus shifted from neighborhood disparities to correlations between variables. These analyses revealed a low correlation between most variables except temporal.

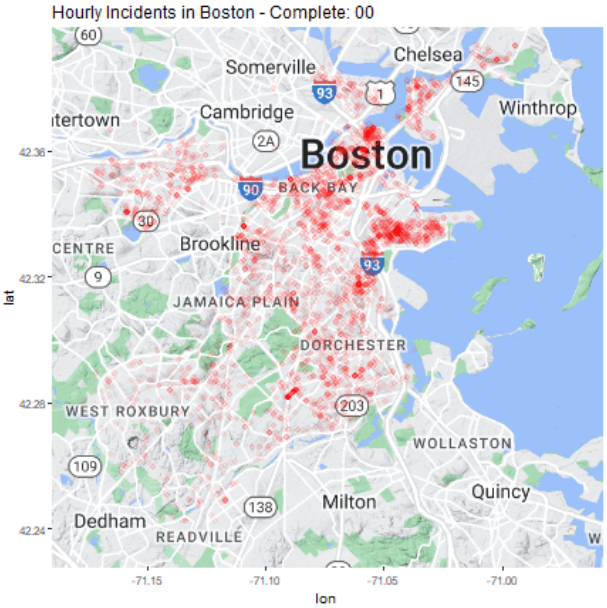
This study explores these key questions:

- Are there areas in Boston that face inequities in non-emergency issues or response times due to specific demographic factors
- Are there non-demographic factors that can predict revealing outcome variables
- What are the residual impacts of these inequities based on annual data reports

Data, Methods, & Statistical Analysis

I. 311 Report Modifications

The City of Boston 311 Service Request dataset[4] contains details regarding 311 reports made from the beginning of 2020 until approximately November of 2021. The original dataset contained 10 columns and 603,632 data instances. Of the 10 attributes, the focus was on "CASE_ENQUIRY_ID", "OPEN_DT", "CLOSED_DT", "TYPE", "CLOSURE_REASON", "longitude", and "latitude". Plotting the reports according to their longitude and latitude revealed several gaps in the data collection process: all of the longitude and latitude values were expected to be within 0.2 of -71.1 and 42.3 respectively. Instances that did not fit these parameters were checked for swapped longitude and latitude values. If these still did not produce the expected values, they were removed from the working dataset. To aid with future analysis, all rows with missing values were removed. Then "CLOSURE_REASON" attribute was taken into consideration. Any rows containing the phrase "duplicate" (case insensitive) were removed. Multiple people reporting on a singular issue as well as form submission errors resulted in several cases being overrepresented.



4 "311 Service Requests - Analyze Boston," n.d., https://data.boston.gov/data

After these changes, the final working dataset had 596,558 individual reports.

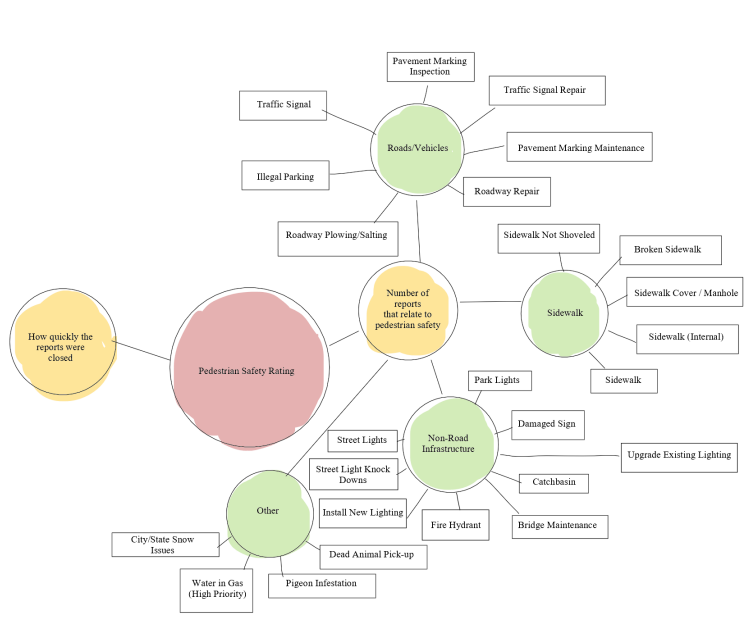
Once the working dataset was determined, the attributes that detailed longitude, latitude, and time of day the case was opened were used to better understand the pulse of the city. The graph below shows cases reported for every hour of the day with 0 being midnight and 23 being 11 P.M. The incident rate begins to pick up at hour 7 and starts to slow down at hour 17. The fewest incidents were reported between hours two and four. Unlike areas with constant bustle that may have a steady stream of incidents at all hours, the graph shows that Boston residents are more likely to follow a particular schedule and are less active or less likely to report incidents after approximately hour 17.

While the dataset itself only had 10 attributes, the reports provided valuable insight regarding the specific non-emergency concerns within Boston. The attribute “TYPE” has 88 unique values which describe the category of the incident being reported. They ranged from “Illegal Parking” to “MBTA Request” to “New Tree Requests” and were contextually supplemented by the “BODY” attribute, which is one of the few attributes that did not have a drop-down list of options to choose from. As a means of visualizing data better and to compensate for the smaller selection of attributes, variable construction was an integral part of the analysis process.

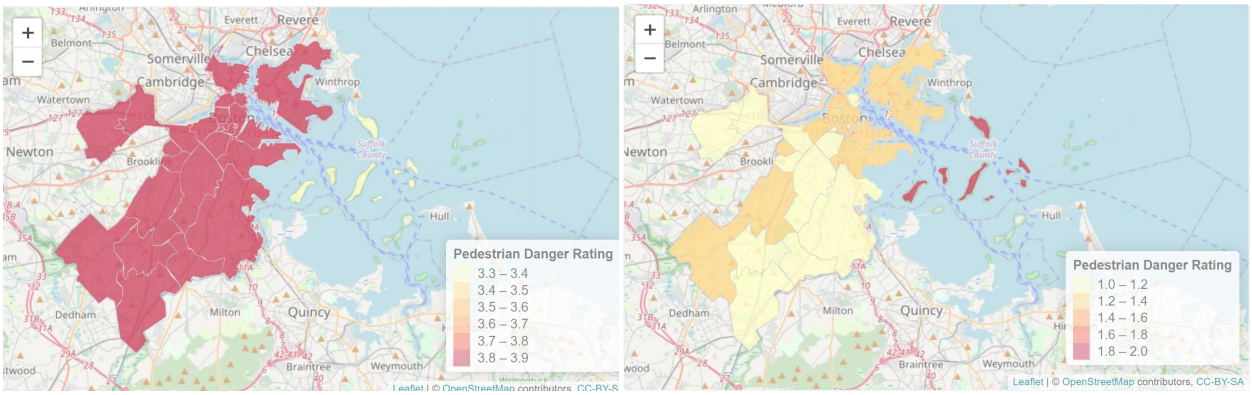
In addition to the creation of a latent variable that assigns Pedestrian Danger Ratings to the reports, there were many temporal variables that were created to better understand the circumstances surrounding the creation, duration, and resolution of each report.

II. Latent Variable Construction

While the “generalized_category” variable was helpful for basic plotting, the consistent proportions throughout the city kept the variable from revealing useful information regarding disparities throughout the community. Pedestrian Danger Rating (PDR) was a latent variable that was constructed to aid analysis. Latent variables are factors that cannot be directly measured or observed. PDR was created by first going through every unique report “TYPE” and selecting all of the categories that pertain to pedestrians. The three main categories are “Road/Vehicle”, “Sidewalk”, and



“Non-Road Infrastructure”. The final category is Other which contains difficult-to-classify types: City/State Snow Issues, Water in Gas (High Priority), Pigeon Infestation, and Dead Animal Pick-up. While these are not traditional pedestrian issues, they have the potential to impact the general safety of those in the vicinity of the issue. Once the issues were determined, each issue was given a value from one through five with one being the least dangerous and five being the most dangerous for pedestrians. The initial reasoning behind this variable was that a difference in PDR between neighborhoods could show a correlation between PDR and other attributes of a neighborhood.



This was done by surveying various people and having them select and rate each report type that pertains to pedestrian safety. The mode of each rating was then assigned as that specific type’s danger rating. The first graph shows the Pedestrian Danger Ratings from 3.3 to 3.9. This was done by taking the average PDR for each neighborhood and plotting it over a map of Boston. The issue with this graph is immediately obvious: The PDR range is 1-5, but all of the values are within 0.6 of each other. While this went against preliminary expectations, it was in line with the exploratory analysis which revealed that there were almost equal proportions of each report’s general category across all of the neighborhoods. The second graph takes the average neighborhood PDR based on both zero and nonzero values. There is slightly more variation in this chart, but again, the values are too close together and geographically inconsistent to interpret any clear patterns throughout mainland Boston. Both charts, however, do have one consistency: while mainland Boston is on one end of the graphed spectrum, the islands are on the opposite end. The differences could be attributed to a variety of factors such as the overall number of reports (a small number can cause skew), the severity of the danger ratings, and a lack of issues typically found in the more central part of the city, namely parking, larger infrastructure issues, and environmental issues, which could explain why the islands have fewer 0 PDR-value reports to lower the average. There are also fewer residents, a lower population density, and fewer nonresidential vehicles in these areas, which can explain why the islands’ PDR is lower when only considering issues that impact pedestrians.

III. Time of Day & Neighborhood

To better understand the issues and needs of Boston neighborhoods, an aggregate measure was created to observe the "OPEN_DT" variable, which gives the date and time each case was reported. The first step was to generate a new variable called "Time of Day," which assigned each report to one of five time slots based on the time the report was created: early morning (ranging from 5 AM to 9 AM), late morning (ranging from 9 AM to 12 PM), afternoon (ranging from 12 PM to 5 PM), evening (ranging from 5 PM to 9 PM), and night (ranging from 9 PM to 5 AM). These time increments were chosen because they are common ways to divide up the day and give more context regarding what portion of the day each neighborhood is experiencing a certain volume of 311 reports.

Next, an SHP file containing boundary data on 25 Boston neighborhoods[5] was used to determine which neighborhood each report was associated with. The longitude and latitude of each report were taken into consideration while creating the variable, which was then used to create a new table. The new table contains three attributes: neighborhood name, time of day, and a count of the number of reports for that neighborhood during that specific time of day.

The aggregate measure provides insight into the daily activity levels of each neighborhood by analyzing the frequency of 311 reports at different times of the day. The measure allows for comparisons between neighborhoods and identifies patterns in each neighborhood's reporting habits. This information is useful for identifying areas of concern and allocating resources accordingly.

Statistical Analysis

I. Correlation

The initial motivation behind this analysis was to determine if geographic or demographic factors could be used to identify patterns in 311 reports, meaning that geospatial data was crucial to the beginning stages of this report. All analyses leading to this portion however did not indicate any definitive connections between where a report was created and the attributes of the report. The disproportionate focus on location thus far was to uncover demographic correlations that could be evident in the report data. Moving forward, the focus shifted to other attributes in the dataset.

Several variables were created to add perspective to existing attributes and to group values in a way that conveyed information more effectively: The main variables created were sec_of_day, sec_resolved, and distance_from_ctd. The first variable (sec_of_day) divides the day into four equal segments, each six hours long, starting at midnight. Attempting to depict anything related to time proved difficult when dividing by 24 hours, so I sec_of_day was created to depict larger intervals of time. The sections are in reference to the time the case was opened or reported. The

5 "Boston Neighborhoods - Analyze Boston," n.d., https://data.boston.gov/dataset/boston-neighborhoods.

following variable (sec_resolved) is conceptually similar, but instead of reflecting the time the case was opened, this variable describes when it was closed. The third variable (distance_from_ctd) was calculated by taking the longitude and latitude of each reported case, and finding out its distance from the center coordinates of Downtown Boston (-71.057083 N, 42.361145 W). Lastly, I created 4 new variables to showcase opening and closing times with respective months and hours for further personal analysis.

The correlation matrix contained the variables section of the day, section resolved, case duration, longitude, latitude, distance from the center of Downtown, the month and hour a case was opened or closed, and the pedestrian danger rating:

	sec_of_day	resolved	case_duration	longitude	latitude	Distance_from_CDT	op_month	op_hour	cl_month	cl_hour	Pedestrian_Danger_Rating
sec_of_day	1.000000000	0.199632522	0.02456182	0.02088433	-0.004229119	0.01160110	-0.011021798	0.940861046	-0.013544848	0.178312788	0.055020363
resolved	0.199632522	1.000000000	-0.02471500	0.03916210	0.054645338	-0.05417822	0.006359262	0.188474135	0.003477463	0.945052870	0.041398859
case_duration	0.024561818	-0.024714995	1.00000000	-0.04209579	-0.028075232	0.04225558	0.032844614	0.026287366	-0.031640755	-0.020509117	-0.062094299
longitude	0.020884330	0.039162099	-0.04209579	1.00000000	0.394835782	-0.64756516	0.014312625	0.016003948	0.018213475	0.039761526	0.067735174
latitude	-0.004229119	0.054645338	-0.02807523	0.39483578	1.000000000	-0.81310315	0.022064427	-0.008857374	0.022153129	0.057065807	0.061923545
Distance_from_CDT	0.011601096	-0.054178217	0.04225558	-0.64756516	-0.813103148	1.00000000	-0.036004584	0.016873251	-0.038485465	-0.055824685	-0.012296060
op_month	-0.011021798	0.006359262	0.03284461	0.01431262	0.022064427	-0.03600458	1.000000000	-0.013017525	0.945726019	0.004404508	-0.003509079
op_hour	0.940861046	0.188474135	0.02628737	0.01600395	-0.008857374	0.01687325	-0.013017525	1.000000000	-0.015581345	0.179031066	0.054928318
cl_month	-0.013544848	0.003477463	-0.03164075	0.01821347	0.022153129	-0.03848547	0.945726019	-0.015581345	1.000000000	0.001026788	-0.007677372
cl_hour	0.178312788	0.945052870	-0.02050912	0.03976153	0.057065807	-0.05582468	0.004404508	0.179031066	0.001026788	1.000000000	0.046095689
Pedestrian_Danger_Rating	0.055020363	0.041398859	-0.06209430	0.06773517	0.061923545	-0.01229606	-0.003509079	0.054928318	-0.007677372	0.046095689	1.000000000

As seen by the correlation matrix above, there were few variables that could be considered correlated. Higher correlations were seen between obvious variables such as the section of day a case was resolved and the hour a case was resolved. Aside from those, the most prominent correlations were between the hour a case was resolved and the month a case was opened with a coefficient of 0.179, the distance from the center of Downtown and hour a case was closed (-0.056), and the month a case was closed and case duration (-0.032).

II. ANOVA and T-Test

Following the correlation matrix, several ANOVAs were conducted. To further confirm that neighborhood had little to no impact on case statistics, an ANOVA was conducted on neighborhood vs case duration:

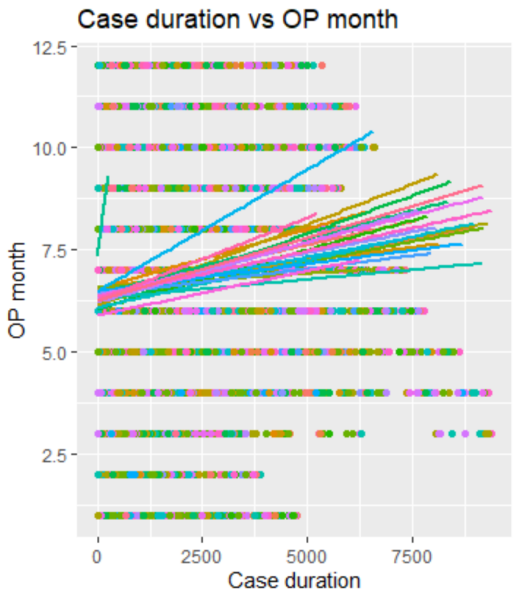
```
                Df    Sum Sq Mean Sq F value Pr(>F)
neighborhood    25 4.781e+06  191226    0.79   0.76
Residuals   270308 6.546e+10  242185
```

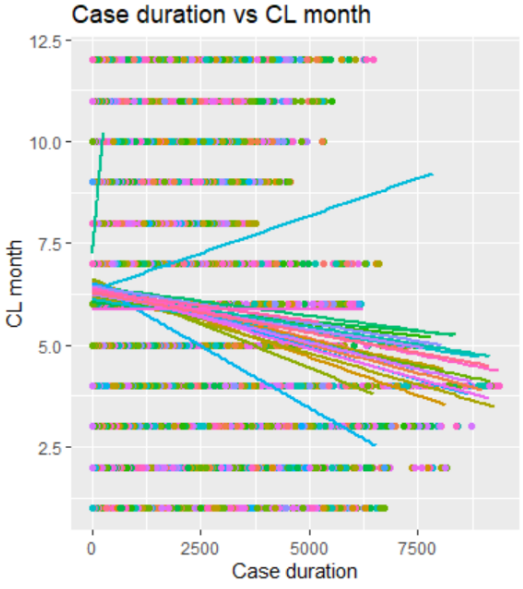
Here, the low F-value (0.79) suggests that there is a slight difference between the group means. The high Pr(>F) indicates that the probability of obtaining this F-value by chance is fairly high, making this statistically insignificant with a low level of confidence. Conversely, further analysis of case closure data yielded different results:

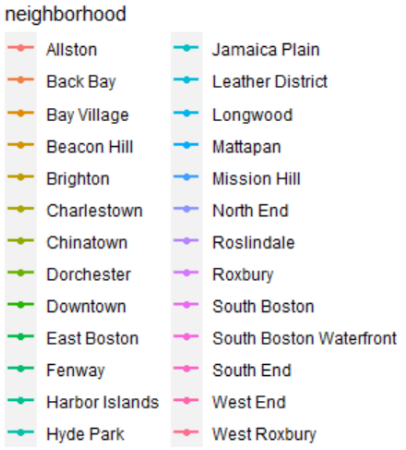
```
                Df    Sum Sq  Mean Sq F value Pr(>F)
cl_hour          1 2.754e+07 27537997   113.8 <2e-16 ***
Residuals   270332 6.544e+10   242080
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                Df    Sum Sq  Mean Sq F value Pr(>F)
cl_month         1 6.554e+07 65543874   270.9 <2e-16 ***
Residuals   270332 6.540e+10   241939
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

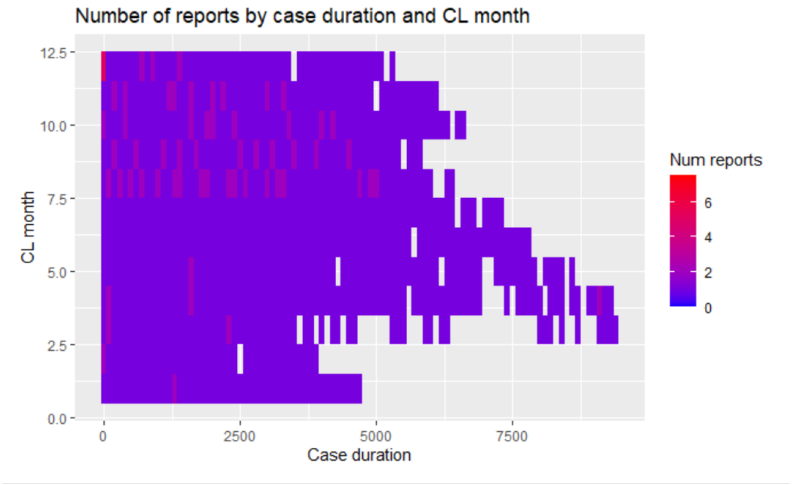
The above ANOVAs were done between case closure hour and case duration and case closure month and duration. Both had fairly high F-values (113.8 and 270.9 respectively) and low Pr(>F) values ($2*e^{-16}$ for both). The high F-value suggests that there is a significant difference between the group means. The high Pr(>F) indicates that the probability of obtaining this F-value by chance is fairly low, making this statistically significant with a high level of confidence.

The first visualization below plots the case duration against the month that cases were opened and closed in. Each color represents a different neighborhood in Boston. The graph shows that case duration greatly differs based on which month the case was closed. Cases are resolved quickest in the month of August and case duration gradually increases until the month of April, which is when cases take the longest to resolve compared to any other month. The ANOVA results paired with the gradual increase depicted in the visual support that there are factors that contribute to different months having different lengths of case durations. Both opening and closing months are shown, but the main focus of the report will be the closure month due to its higher correlation coefficient.





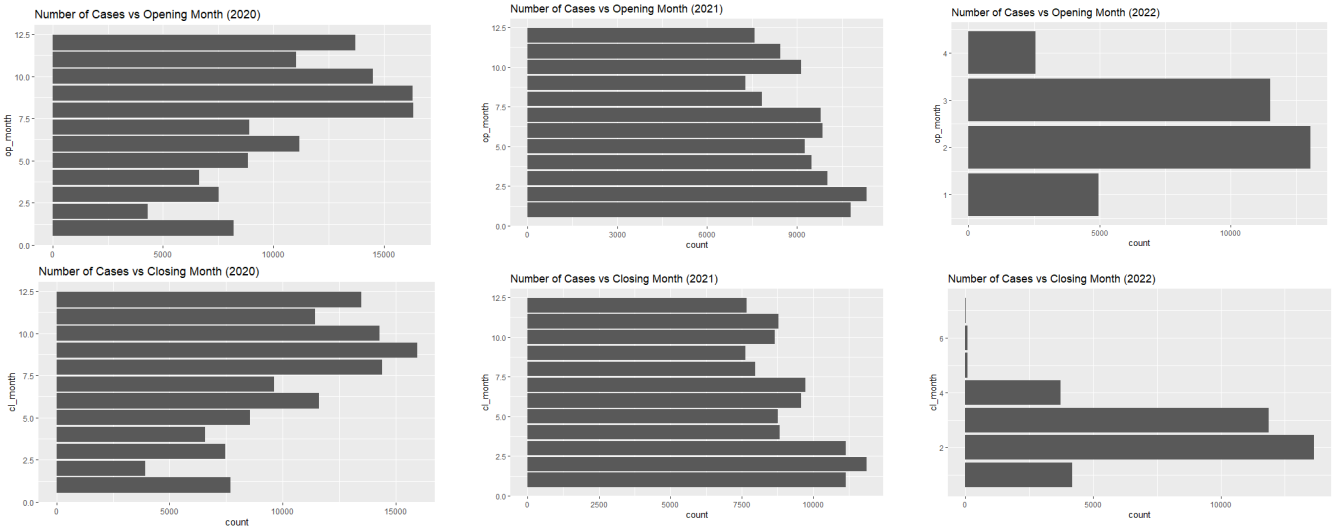




The second visualization conveys similar information to the graph above but is color-coded by the density of reports that are grouped by case duration. It can be seen that aside from a select few instances in the later half of the year, there are not many cases that share duration. At most, 7 cases during any given closure month will share case duration, making it fairly insignificant among hundreds of thousands of reports.

III. Further Investigation - Opening and Closing Month

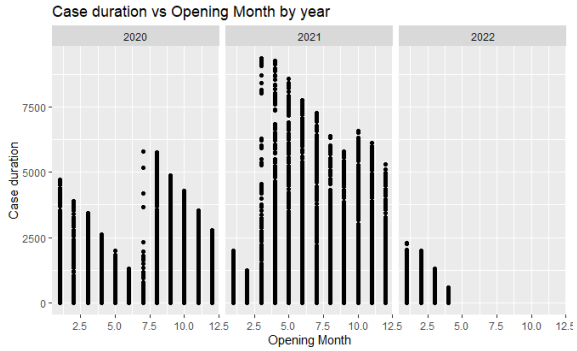
The lengthier case duration during the early months of the year can likely be attributed to weather-related restrictions. Snowstorms, cold weather, and limited sunlight can impact how frequently roads and infrastructure can be repaired, and consequently, how long cases stay open. To gain a deeper understanding, however, it is important to look at the data with respect to the year. Below are a series of charts that show the number of reports filed or completed within a given month by year:

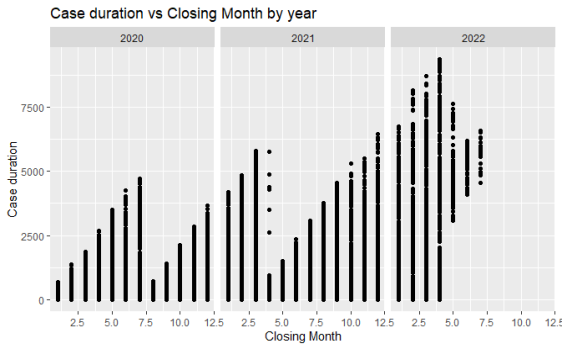


Throughout 2020, cases seem to open and close fairly consistently and in a timely manner. However, in 2021, slight differences emerge: there are more cases opened starting in April than there are cases being closed even when considering the slight difference in axis range. The charts

for 2022 look vastly different from one another but that can be attributed to incomplete data for that year.

Plotting against case duration instead of the number of reports gives more context regarding response trends:





The first set of plots shows the month cases were opened along with how long it takes to resolve the case. The second set of plots shows the month cases were closed along with how long it took to resolve the case. Both graphs have clear patterns: the first set has a peak that tapers down, suddenly shoots up slightly higher, and tapers down again. The pattern continues into the following year as well. The second set has a similar pattern but the duration steadily increases before shooting down instead of decreasing like the previous set. Cases created in June had a lower all-around case duration in 2021 with a sharp uptick on the upper bound of the case duration range in August where the cases once again start to get resolved consistently quicker before peaking once again in April of 2021. The graphs of the number of cases reported by month do not have any patterns that resemble the charts above, which likely means that case duration is impacted by external factors and not limited by the number of cases available. Low upper bounds for case duration can be attributed to several factors: fewer cases may be resolved during these months so there might be fewer chances for cases to take longer to resolve or cases may be resolved quickly due to more employees being present (there could be more workers on vacation or there may be additional workers available during school breaks). There may be more instances of quickly resolved cases during these months as well as fewer instances of long-term cases being resolved (more parking violations and fewer pothole requests). An important factor to note, however, is that with each peak, the upper bound of the case duration grows higher than it has before. A more complete dataset with reports from the start of the system, August 2015, should be examined to see if this is a long-term pattern. One possible explanation for this pattern is that case reports and duration may have been impacted by waves of COVID-19. Fewer issues could have been spotted, resources can experience less wear and tear, and worker availability may be restricted from time to time. The peaks may also be a result of extremely non-emergency requests being fulfilled when resources are available. Issues such as “New Tree Request” do not require immediate attention and can be put off until the weather is warmer.

IV. Conclusion and Concerns

The discovery of a pattern in the case duration range brings forth several questions. Higher case durations in a given set of months do not imply a change in productivity, but there was hesitancy in the preliminary stages of constructing this variable due to the possibility of reporting bias. Reporting bias could be caused by a lack of knowledge surrounding the 311 system, varying standards for what constitutes a report, and more.

An inconsistent and patterned case duration range also suggests that 311 reports are not streamlined for efficiency. Either issues are not being addressed or reported in a timely manner or preventative measures are not made to facilitate a more consistent and streamlined process. For example, if potholes cannot be filled during certain months, roads can be updated more frequently to reduce the chances of potholes, a special process could be developed to address potholes in the winter months, etc. By doing this, the city can keep the number of cases reported proportional to the labor they have to address these specific issues and possibly even keep case numbers consistent throughout the year. Ensuring that issues are addressed in a timely manner helps with neighborhood upkeep and also helps residents maintain faith in the reporting system. By being aware of these trends, the city of Boston can prepare by hiring staff in accordance with their needs.

If the increased case durations in certain months are due to backlog from other months, then addressing this trend is imperative. Important cases can be pushed off if it is more time-consuming than other tasks or if it gets temporarily lost in the chaos of a case backlog. Failing to address important needs, while important in all communities, is especially important to areas with disproportionately higher populations of vulnerable groups. Prolonged unresolved issues can cause unintended harm to specific communities and have long-lasting impacts compared to their less vulnerable counterparts.

V. Future Analysis Steps:
    - Include data from 2015 onward to observe long-term trends
    - Perform clustering on month and duration variables
    - Include more demographic data such as population density, percentage of students within the population at any given time of the year, traffic data, infrastructure data, and more
    - Include financial data to understand the distribution of resources for every community
    - A more detailed breakdown of Boston (blocks instead of neighborhoods)
    - NLP for the expanded report reason (short answer response)

VI. Concerns:

- As mentioned before, this dataset does not have enough years of data to fully understand long-term trends in the reporting system. Adding data from more years can give us insight into more normal happenings that were not impacted by the COVID-19 crisis.
- When cleaning, the step that removes all rows with NA values is limiting for various reasons:
    - If optional report fields exist, a portion of reports are being omitted due to a lack of understanding regarding the reporting system and not for error or outlier purposes
    - This also means that any cases that were still unresolved in July 2022 were removed from the working dataset. At the very least, unfinished cases should be separately studied to uncover any potential patterns and causes

Bibliography

"311 Service Requests - Analyze Boston," n.d. https://data.boston.gov/dataset/311-service-requests.

"Boston Neighborhoods - Analyze Boston," n.d. https://data.boston.gov/dataset/boston-neighborhoods.

SAMHSA - the Substance Abuse Mental Health Services Administration. "Greater Impact: How Disasters Affect People of Low Socioeconomic Status." Accessed April 18, 2023. https://www.samhsa.gov/.

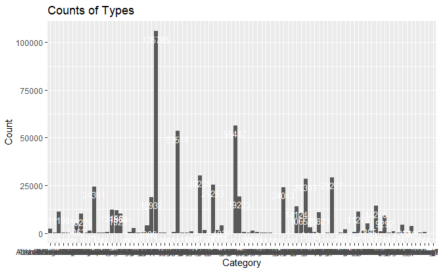
US EPA. "EPA Report Shows Disproportionate Impacts of Climate Change on Socially Vulnerable Populations in the United States | US EPA," September 2, 2021. https://www.epa.gov/newsreleases/epa-report-shows-disproportionate-impacts-climate-change-socially-vulnerable.

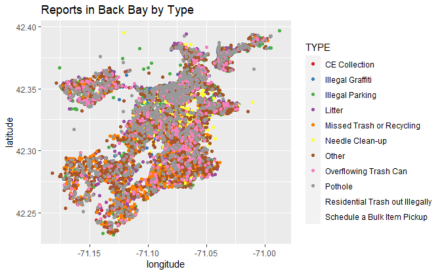
World Health Organization: WHO. "Urban Health." *Www.Who.Int*, October 29, 2021. https://www.who.int/news-room/fact-sheets/detail/urban-health.

Appendix

The following section details the reasoning behind the creation of the manifest variables used in this report.

I. Generalized Type

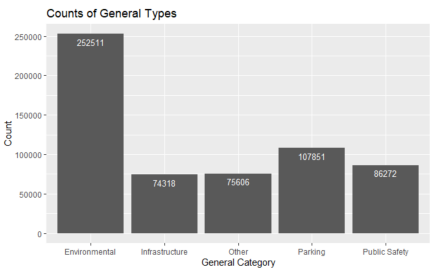


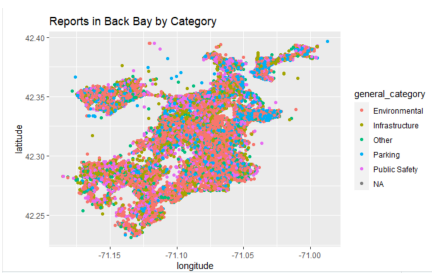


As indicated above, the "TYPE" attribute has many unique values, making it difficult to visualize in a standard graph or map. The graphs below show the distribution of types and the location of each report by type To rectify this, a new variable called "general_category" was constructed where each of the 88 categories was regrouped into one of five categories: environmental, infrastructure, public safety, parking, and other. While there is a disproportionate number of "TYPE"s among each category (infrastructure has a broader range of subtypes than parking and other) approximately half of the reports fall under the environmental category, indicating a deficiency in the city of Boston's waste management practices.

Any reports pertaining to parking violations were categorized as parking, any reports involving recycling, trash, or general waste management as well as tree-related violations were categorized as environmental, Any issues regarding public roads and structures were classified as infrastructure, and any issues that can prove to be a danger to the general public was categorized as a public safety issue. Report types that did not fall into any of the four categories were classified as other.

When mapping out the reports in accordance with their general types, the proportion of each category was fairly consistent throughout all of the Boston neighborhoods. The charts below show the distribution of general categories and the location of each report by general category, and the smaller selection of variables made the graphs easier to decipher. Though there aren't many patterns that can be seen throughout the map, it can be noted that parking violations are represented at a higher rate





in the northern half of the city. This is likely due to the more intense parking restrictions in this part of Boston. Neighborhoods such as Allston and Brighton have more parking options that are not metered and are available to nonresidential vehicles, making for fewer violations in those areas.

II. Code Excerpts

Loading & Cleaning the 311 Report Data

```r
```{r 311}
#Read in CSV data
df<-read.csv('BOS311.csv')

#create df that has important columns
mod_df<-df[c('CASE_ENQUIRY_ID', 'OPEN_DT', 'CLOSED_DT', 'TYPE', 'CLOSURE_REASON',
'DEVICE', 'latitude', 'longitude')]
NA_df <- mod_df[rowSums(is.na(mod_df)) > 0,]
cleaning <- na.omit(mod_df)
#cleaning now contains all of rows of mod_df that aren't missing long/lat values
range(cleaning$latitude)
for (i in 1:nrow(cleaning)) {
  if (cleaning[i, "latitude"] < 0 || cleaning[i, "longitude"] > 0) {
    tmp <- cleaning[i, "latitude"]
    cleaning[i, "latitude"] <- cleaning[i, "longitude"]
    cleaning[i, "longitude"] <- tmp
  }
}
mean_col1 <- mean(cleaning$latitude)
mean_col2 <- mean(cleaning$longitude)

cleaning <- cleaning[abs(cleaning$latitude - mean_col1) <= 2 &
abs(cleaning$longitude - mean_col2) <= 2, ]
cleaning1 <- cleaning %>% sample_n(100000)

# Convert OPEN_DT and CLOSED_DT to date-time objects
cleaning$OPEN_DT <- as.POSIXct(cleaning$OPEN_DT, format = "%Y-%m-%d %H:%M")
cleaning$CLOSED_DT <- as.POSIXct(cleaning$CLOSED_DT, format = "%Y-%m-%d %H:%M")

#removes N/A values
cleaning <- cleaning[!is.na(cleaning$TYPE) & cleaning$TYPE != "", ]
```
```

```
# read in the Boston neighborhood shapefile to get which Neighborhood each case
is in
neighborhoods <- st_read("boston_neighborhoods.shp")
my_data <- cleaning[, c("longitude", "latitude")]
my_points <- st_as_sf(my_data, coords = c("longitude", "latitude"), crs = 4326)
my_points <- st_transform(my_points, st_crs(neighborhoods))
my_points_with_neighborhoods <- st_join(my_points, neighborhoods)
neighborhood_names <- my_points_with_neighborhoods$Name
length(neighborhood_names)
neighborhood_names <- neighborhood_names[!is.na(neighborhood_names)]
neighborhood_names <- neighborhood_names[1:nrow(cleaning)]
cleaning$neighborhood <- neighborhood_names

#read in as a different variable in case I want to alter anything
boston_neighborhoods <- st_read("Boston_neighborhoods.shp")
```

Establishing Manifest Variables

I. Case Duration

```
```{r duration}
# Calculate case duration
cleaning$case_duration <- as.numeric(difftime(cleaning$CLOSED_DT,
cleaning$OPEN_DT, units = "hours"))
```
```

Establishing Latent Variables

I. Generalized Type

```
```{r new: general_category}

# create a vector of the types you want to map to each general category
enviro_types <- c("Residential Trash out Illegally", "Litter", "Needle Clean-up",
"Illegal Dumping", "Trash on Vacant Lot", "Recycling Cart", "Recycling Cart
Return", "Overflowing Trash Can", "Litter Basket Maintenance", "CE Collection",
"Missed Trash or Recycling", "Schedule a Bulk Item Pickup", "Dead Animal
Pick-up", "Overflowing/Unkept Dumpster", "Litter Basket", "Construction Debris")
infra_types <- c("Major System Failure", "Traffic Signal", "Pavement Marking
Inspection", "Roadway Plowing/Salting", "Broken Sidewalk", "Catchbasin", "Traffic
```
```

```
Signal Repair", "Sidewalk Cover / Manhole", "Utility Casting Repair", "Knockdown
Replacement", "BWSC Pothole", "Bridge Maintenance", "Traffic Signal Studies",
"Upgrade Existing Lighting", "Abandoned Building", "General Traffic Engineering
Request", "Park Improvements", "Roadway Repair", "Street Light Knock Downs",
"Damaged Sign", "MBTA Request", "General Lighting Request", "Park Lights",
"Pothole", "Sidewalk", "Pavement Marking Maintenance", "New Sign, Crosswalk or
Marking", "Planting")
pub_safety_types <- c("Install New Lighting", "Street Lights", "Illegal
Graffiti", "Abandoned Vehicle", "Broken Park Equipment", "Dead Tree Removal",
"Abandoned Bicycle", "New Sign", "Crosswalk or Marking", "Fire Hydrant", "News
Boxes", "Bicycle Issues", "Tree in Park", "Water in Gas (High Priority)", "Fire
Department Request", "City/State Snow Issues", "Sidewalk (Internal)", "Sidewalk
Not Shoveled", "Tree Emergencies", "New Tree Requests", "Tree Pruning", "Short
Measure - Gas")
parking_types <- c("Illegal Parking", "Space Savers", "Parking Front/Back Yards
(Illegal)", "Parking Meter Repairs", "Valet Parking Problems", "Private Parking
Lot Complaints", "Municipal Parking Lot Complaints", "Short Term Rental")
other_types <- c("Other", "Rodent Sighting", "Pigeon Infestation", "Sticker
Request", "Missing Sign", "Illegal Auto Body Shop", "Scanning Overcharge", "Item
Price Missing", "No/Wrong Gas Price", "Product Short Measure", "Cemetery
Maintenance Request", "Student Move-In Issues", "Unit Pricing Wrong/Missing",
"Illegal Posting of Signs", "Scale Not Visible", "Aircraft Noise Disturbance")

# use ifelse to map each type to its general category
cleaning$general_category <- ifelse(cleaning$TYPE %in% enviro_types,
"Environmental",
                            ifelse(cleaning$TYPE %in% infra_types,
"Infrastructure",
                                   ifelse(cleaning$TYPE %in% pub_safety_types,
"Public Safety",
                                          ifelse(cleaning$TYPE %in%
parking_types, "Parking",
                                                 ifelse(cleaning$TYPE
%in% other_types, "Other", "")))))

```

II. Pedestrian Danger Rating

```{r PDR}
```

```
library(sqldf)

# Define the rating values as a named vector
rating_values <- c("Pavement Marking Maintenance" = 1,
                   "Catchbasin" = 1,
                   "Bridge Maintenance" = 1,
                   "Fire Hydrant" = 1,
                   "Damaged Sign" = 2,
                   "Park Lights" = 2,
                   "Upgrade Existing Lighting" = 2,
                   "Dead Animal Pick-up" = 2,
                   "Pigeon Infestation" = 2,
                   "Install New Lighting" = 2,
                   "Roadway Repair" = 3,
                   "Roadway Plowing/Salting" = 3,
                   "Sidewalk Cover/Manhole" = 3,
                   "Sidewalk (Internal)" = 3,
                   "Sidewalk" = 3,
                   "Pavement Marking Inspection" = 3,
                   "Traffic Signal Repair" = 4,
                   "Traffic Signal" = 4,
                   "Illegal Parking" = 4,
                   "Street Light Knock Downs" = 4,
                   "Water in Gas (High Priority)" = 4,
                   "Sidewalk Not Shoveled" = 5,
                   "Street Lights" = 5,
                   "Broken Sidewalk" = 5,
                   "City/Snow Snow Issues" = 5)

# Use a subquery to join the cleaning data with the rating values
cleaning_with_rating <- sqldf("SELECT cleaning.*,
                              CASE
                                WHEN TYPE = 'Catchbasin' THEN 1
                                WHEN TYPE = 'Bridge Maintenance' THEN 1
                                WHEN TYPE = 'Fire Hydrant' THEN 1
                                WHEN TYPE = 'Damaged Sign' THEN 2
                                WHEN TYPE = 'Park Lights' THEN 2
                                WHEN TYPE = 'Upgrade Existing Lighting' THEN 2
                                WHEN TYPE = 'Dead Animal Pick-up' THEN 2
```

```
                                WHEN TYPE = 'Pigeon Infestation' THEN 2
                                WHEN TYPE = 'Install New Lighting' THEN 2
                                WHEN TYPE = 'Roadway Repair' THEN 3
                                WHEN TYPE = 'Roadway Plowing/Salting' THEN 3
                                WHEN TYPE = 'Sidewalk Cover/Manhole' THEN 3
                                WHEN TYPE = 'Sidewalk (Internal)' THEN 3
                                WHEN TYPE = 'Sidewalk' THEN 3
                                WHEN TYPE = 'Pavement Marking Inspection' THEN 3
                                WHEN TYPE = 'Traffic Signal Repair' THEN 4
                                WHEN TYPE = 'Traffic Signal' THEN 4
                                WHEN TYPE = 'Illegal Parking' THEN 4
                                WHEN TYPE = 'Street Light Knock Downs' THEN 4
                                WHEN TYPE = 'Water in Gas (High Priority)' THEN 4
                                WHEN TYPE = 'Sidewalk Not Shoveled' THEN 5
                                WHEN TYPE = 'Street Lights' THEN 5
                                WHEN TYPE = 'Broken Sidewalk' THEN 5
                                WHEN TYPE = 'City/Snow Snow Issues' THEN 5
                                ELSE 0
                            END AS Pedestrian_Danger_Rating
                            FROM cleaning")
```

III. Correlation Matrix

```{r corr}
#install.packages("GGally")
library(GGally)
vars <- c("sec_of_day", "resolved", "case_duration", "longitude", "latitude",
"Distance_from_CDT", "op_month", "op_hour", "cl_month", "cl_hour",
"Pedestrian_Danger_Rating")

correlations <- cor(df[,vars])
View(head(df))
View(correlations)
ggpairs(df[,vars])

vars <- c("case_duration", "longitude", "latitude", "op_hour", "cl_hour",
"Pedestrian_Danger_Rating")
```

```
correlations <- cor(df[,vars])
View(head(df))
View(correlations)
ggpairs(df[,vars])
```

IV. ANOVA

```
```{r ANOVA_visualization}

fit1 <- aov(case_duration ~ cl_month, data = df)
summary(fit1)
fit2 <- aov(case_duration ~ cl_hour, data = df)
summary(fit2)

ggplot(data = df, aes(x = case_duration, y = cl_month, color = neighborhood)) +
geom_point() +
geom_smooth(method = "lm", se = FALSE) +
labs(title = "Case duration vs CL month", x = "Case duration", y = "CL month")
````
```

V. Case Duration vs Closing Month by Year

```
```{r graphs}
library(ggplot2)

# Extract year from CLOSED_DT variable
df$clyear <- lubridate::year(df$CLOSED_DT)

# Graph case_duration vs cl_hour for each year
ggplot(df, aes(x = cl_month, y = case_duration)) +
  geom_point() +
  facet_wrap(~ clyear, ncol = 3) +
  labs(title = "Case duration vs Closing Month by year", x = "Closing Month", y =
"Case duration")
```
```