Deliverable 2 Report

Problem Statement

There are three main issues to tackle with this project:

1. Creating a historical database of the City of Boston's 311 program.
2. Ensuring the service is efficient.
3. Ensuring the service is equitable.

Data Cleaning and Collection Steps

Data was downloaded from the following source link, this contained datasets from 2011-2023. Luckily, the column headers for all the years were consistent, so we did not need to do any modifications before merging the data. We used the merged dataset as the starter dataset for the upcoming cleaning steps.

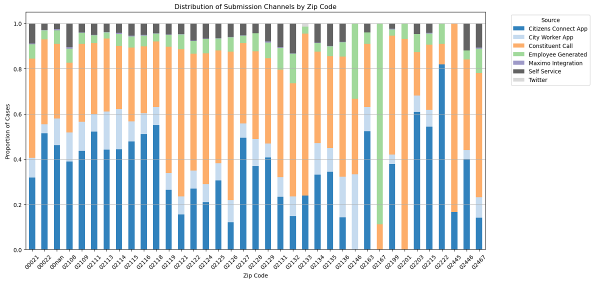
Cleaning was the more rigorous process since we had a total of 2.7M rows, and 1.4M individual null values. The picture columns were dropped as they contained a total of 4.2M null values and we decided not to use them. The SLA_target_date column had 500k empty values, and thousands of unrealistic values for an SLA (on the order of 10^4) related to the completion of tasks associated with the 311 service. Additionally, we have not found a use for this column in any of our base questions, and our prospective extension, so we will be dropping it for now. Case title is redundant, since we have subject/reason and other columns describing the issue type, and we will drop it especially since it has a few hundred empty values. The on_time column has some issues, since it works together with target SLA, and we will not be using it in our analysis hence we will drop it as well. For ["fire_district", "pwd_district", "police_district", "ward", "precinct"] are columns we will not use for base or extension projects. City council and neighborhood services district were labeled -1 to fill missing values. Closure reason was summarized to 1 word matching the provided dictionary. And a reverse Geocoding API was used to fill 600k location zip codes, using their latitude and longitude. Although we still had Nan values at this point, we at least narrowed the scope of the data that we used. Closed_dt, resolution time had 195k empty values, but were important to the final base question, and longitude and latitude had 44k empty values, this was a decent amount to drop so we were conservative with these columns specifically and kept the rows.

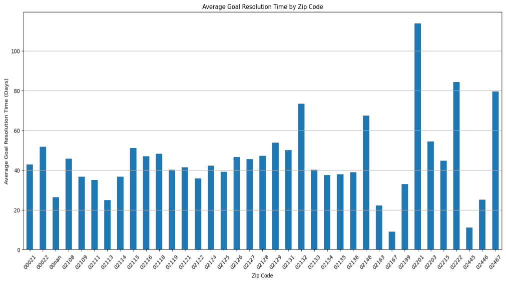
Finally, since we have defined the processing pipeline for data entries, we were able to write a script that periodically (daily- matching the update rate of the website) scrapes the 311 dataset, and fetches the most recent entries. The newly fetched data is appended to the bottom of a PowerBI push dataset, powering an online PowerBI report that can be accessed using this link. We believed this was an elegant supplementary solution to the first problem statement and decided to proceed with its implementation during our data processing, collection and cleaning phase.

Exploratory Data Analysis

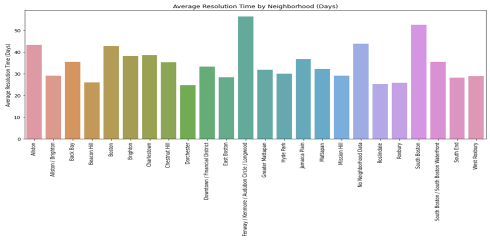
The exploratory data analysis aims to uncover trends and patterns in the utilization of the 311 services. These trends and patterns answer some questions and helps to ask new ones. Observing trends between some of the neighborhoods and their resolution times lead us to notice the correlation between a variety of factors, such as some sources handling a greater number of cases, certain neighborhoods having a faster resolution time, etc.

For example some observations/patterns we made are visible in these 2 bar graphs. Through the analysis of the first bar graph, we observe the differences in the channel source by each zip code and the second graph can help to draw some meaningful conclusions. As we can see, most zip codes have a somewhat identical distribution of sources used to file a service request. To see what this means we can take a look at the average goal resolution time by zip code. From this, we can see that certain zip codes the majority of their submission channel is a constituent call take significantly longer than others, or that there is a zip code in which the majority of the distribution of the submission channels is employee generated and that it has the lowest goal resolution time. The patterns and observations could lead to systemic inefficiencies or resource allocation disparities that need to be addressed. The shift towards digital submission channels underscores the need to focus on digital inclusivity





Similarly, we noticed that the distribution of the average resolution time between neighborhoods was not uniform. This could be due to discrimination such as more high income areas having lower resolution times, or due to circumstance such as only one type of problem that requires more work occurring in specific neighborhoods. We noted this trend and as a result, it influenced our bas and extension questions.



Visualizations

Base Question 1: What is the total volume of requests per year, or how many 311 requests is the city receiving per year?

Plot can be seen on page 1 of the PowerBI link, as well as the first plot in the notebook used to answer base questions. The volume of requests started at around 60K in 2011, and steadily increased until 2017, and has stagnated around the range (250K-270K) since then.

Base Question 2: Which service requests are most common for the city overall AND by NEIGHBORHOOD and how is this changing year over year by SUBJECT (department), REASON,QUEUE?

REASON:

Enforcement & Abandoned Vehicles was the most common REASON for the city overall coming in at around 398K total requests over the 2011-2023 period. In the earlier years, there were changes in the most requested (2011-2012) Sanitation was the most common reason, (2013 and 2015) it was Street Cleaning, (2014) was Highway Maintenance, but the remaining years were dominated by Enforcement & Abandoned Vehicles.

For individual neighborhoods, we recommend using PowerBI, or the notebook for exploration as fulfilling the above requirement for each one of the cities would crowd the report, but generally we noticed the following to be appearing most often across different neighborhoods [“Enforcement & Abandoned Vehicles”, “Street Cleaning”, “Sanitation”, “Highway Maintenance”, “Code Enforcement”].

SUBJECT:

Public Works Department subject dominated with 1.53M total requests from 2011-2023. It was the most common request SUBJECT for each one of those individual years as well. This dominance was also present in the neighborhood data. Some other subjects of interest include [“Transportation - Traffic Division”, “Inspectional Services”, “Parks & Recreation Department”, “Mayor's 24 Hour Hotline”].

QUEUE:

There are 181 unique queues, with a decent distribution of requests across queue categories. This section is best explored through PowerBI/notebook.

Base Question 3: How is the case volume changing by submission channel SOURCE?

Constituent Call, and Citizens Connect app, are the most common forms of request submission (1.12M, and 0.98M) respectively, other notable means include City Worker App with 0.27M, Employee Generated with 0.16M, and Self-Service with 0.16M, the remaining means are negligible.

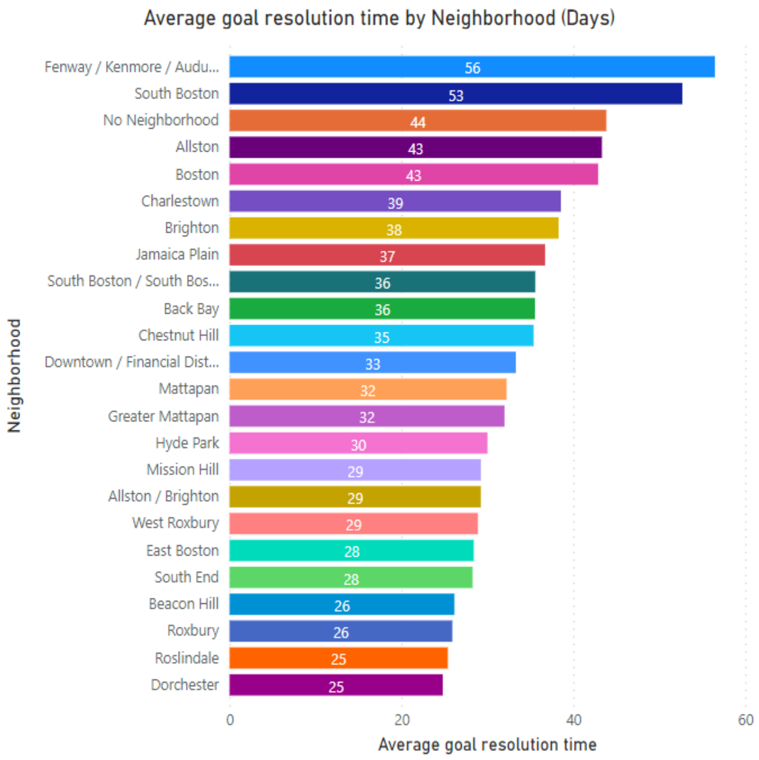
Base Question 4: What is the average # of daily contacts by year?

The average number of daily contacts per year mirrors the total volume of requests, and has evolved from around 160 to stabilizing near 700 daily.

Base Question 5: Volume of top 5 request types

Parking Enforcement had 352K, Requests for Street Cleaning had 187K, Scheduling a Bulk Item Pickup had 164K, Missed Trash/Recycling/Yard Waste/Bulk Item had 137K, and Request for Pothole Repair had 124K.

Base Question 6&7: Average goal resolution time by QUEUE and neighborhood



This is best viewed through PowerBI or through interactive notebook, due to the possible permutations of graphs. However, this question is what prompted our interest in the first proposed extension project.

With raw visualization, and no further data cleaning, Fenway has more than double the time of Dorchester. We thought this entire topic (not just Fenway vs Dorchester) needed more exploration especially since we are hoping to investigate efficiency and equity, and this specific base question tackles both. Combining extra work on this with the social vulnerability index seemed like an excellent branching point for our project that answers both base and extension questions in the spirit of the original task.

Base Question 8: Percent of closed, open, no data case statuses?

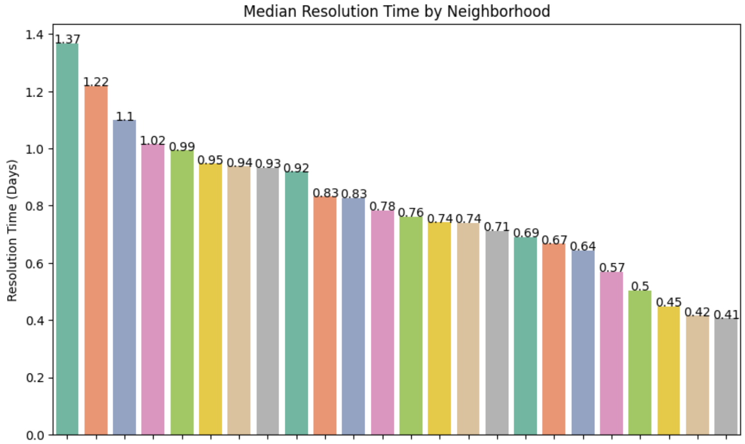
We see around 93% closed, and 7% open cases overall for the entire 2011-2023 period.

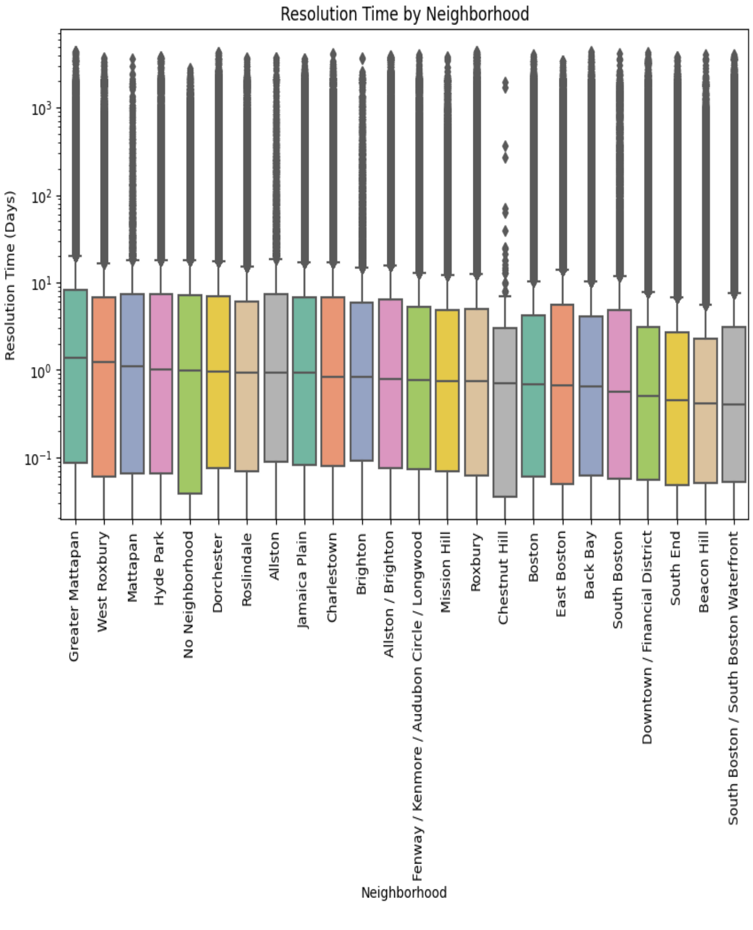
Extension proposal

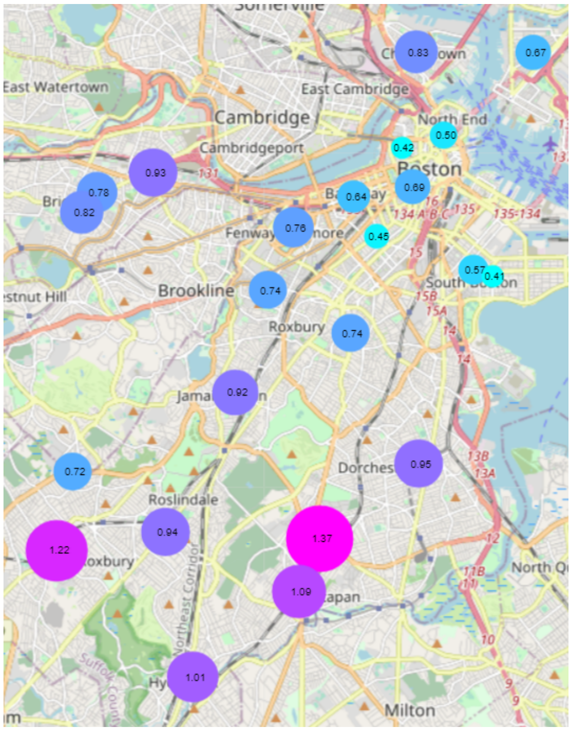
A Better Measure of Resolution Time

Further examination of the resolution time distributions of each neighborhood reveals resolution times in the thousands of days. Looking below we see that resolution time tends to exist mostly in the range of a few hours to a few days, with some cases where resolution time is on the order of 10s of days. 100s, and 1000s of days is much less prominent, but evidently their presence creates a sizable difference in the mean and median values for each neighborhood. Now, we are looking at resolution times that are much smaller than the average which was on the order of 10s of days.

When looking at the physical locations of these medians, we can see some patterns forming. There may be a correlation between the distance from the center of Boston and the median time of task resolution.





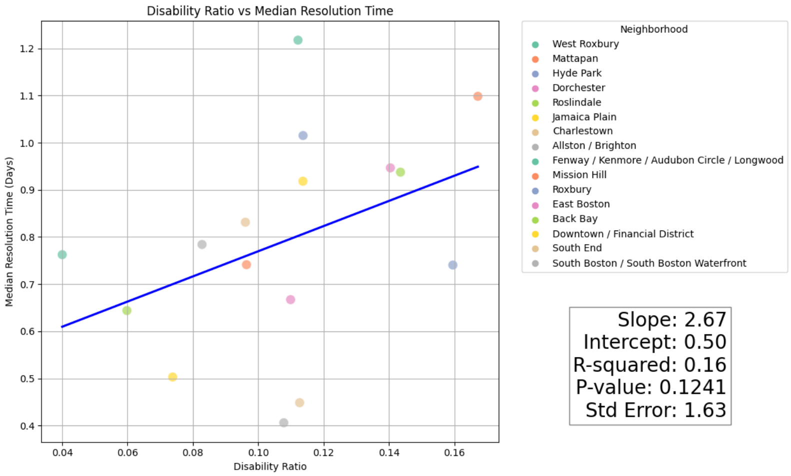


Visualizations (Extension)

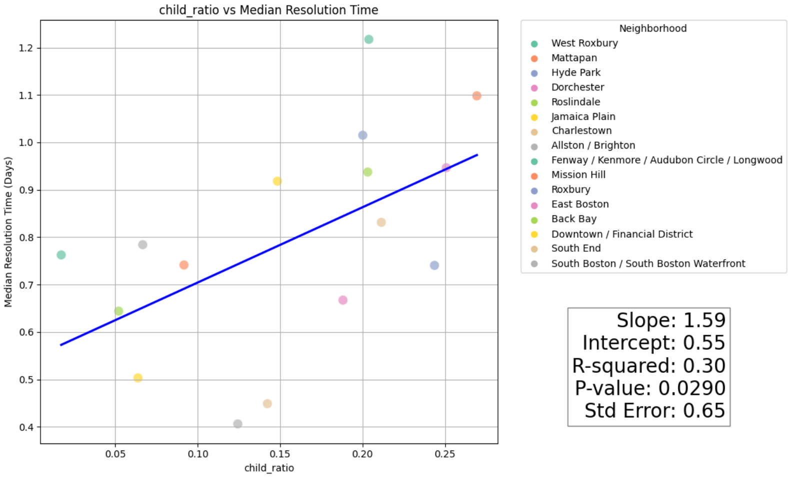
Social Vulnerability Dataset

We integrated the Social Vulnerability dataset for one of our extension projects. We mapped the median resolution time we calculated in the original dataset, and appended it to the Social Vulnerability Index by matching the neighborhoods. This allowed us to directly compare the ratios of certain vulnerability indices to the median resolution time. We originally got counts for different vulnerability measures, but ratios were derived through division by the total number of people in a certain area.

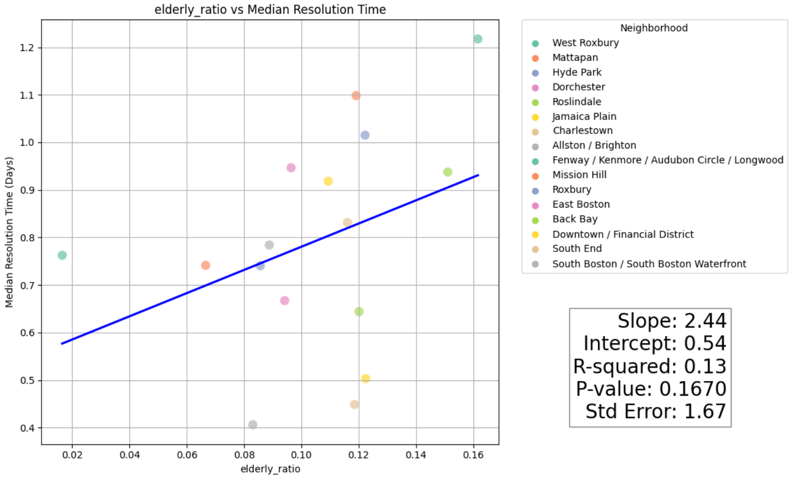
Initial Results



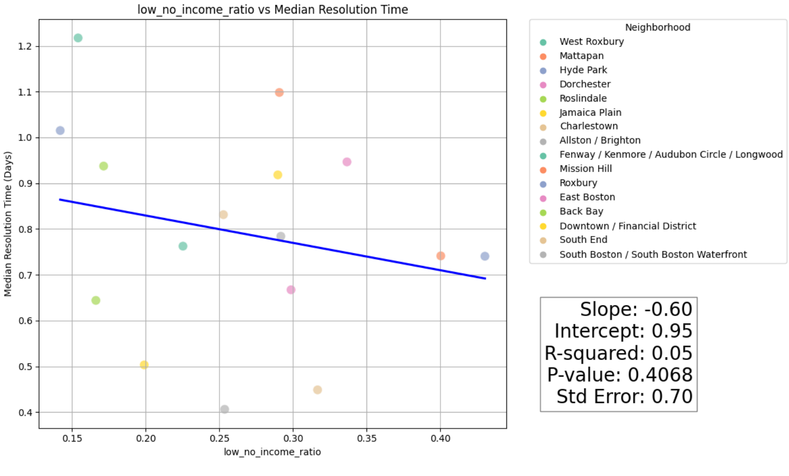
Slope: 2.67
Intercept: 0.50
R-squared: 0.16
P-value: 0.1241
Std Error: 1.63



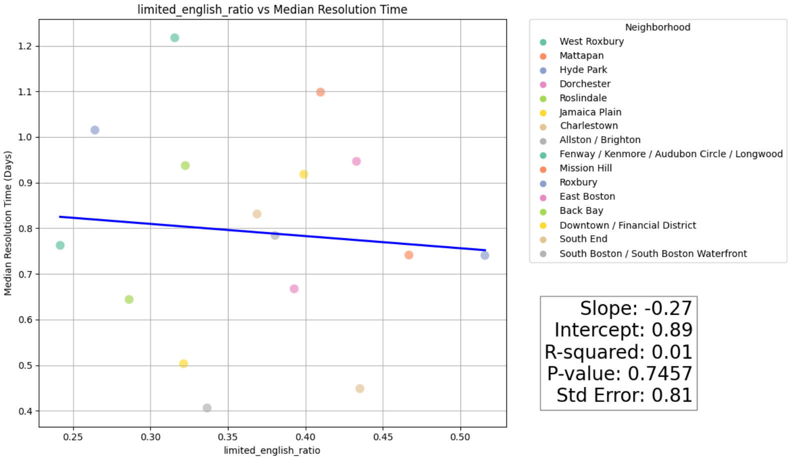
Slope: 1.59
Intercept: 0.55
R-squared: 0.30
P-value: 0.0290
Std Error: 0.65

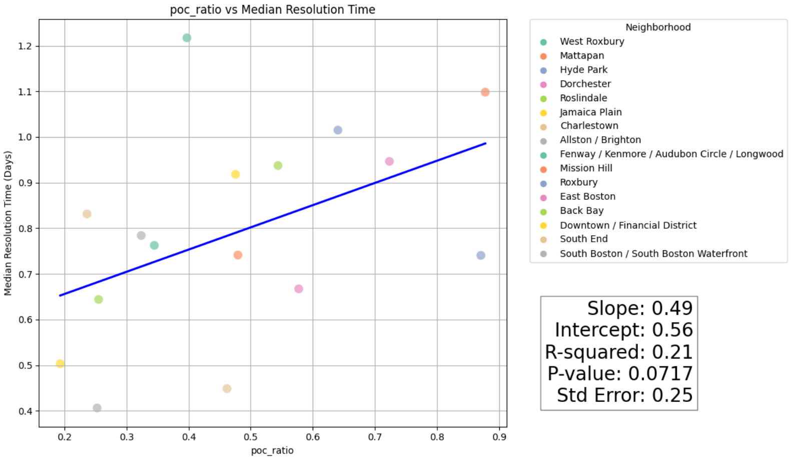


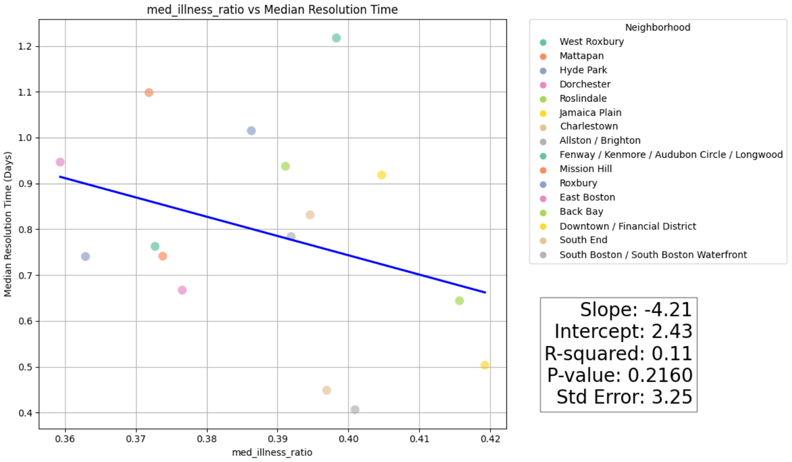
Slope: 2.44
Intercept: 0.54
R-squared: 0.13
P-value: 0.1670
Std Error: 1.67



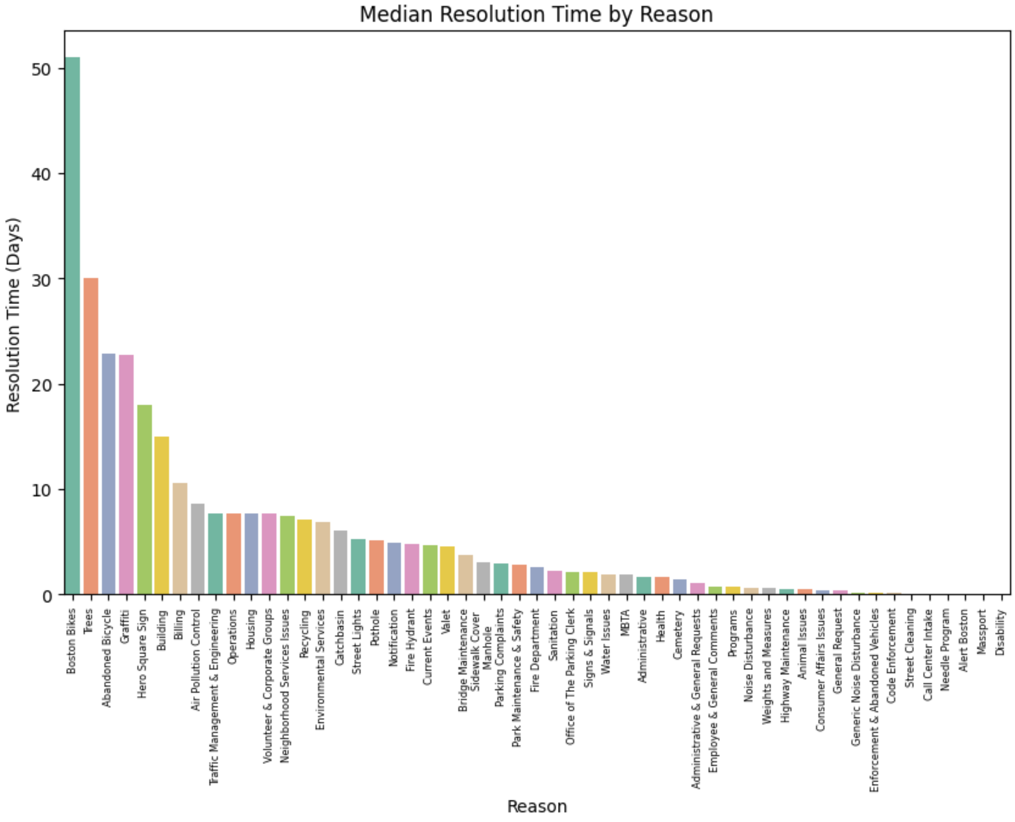
Slope: -0.60
Intercept: 0.95
R-squared: 0.05
P-value: 0.4068
Std Error: 0.70

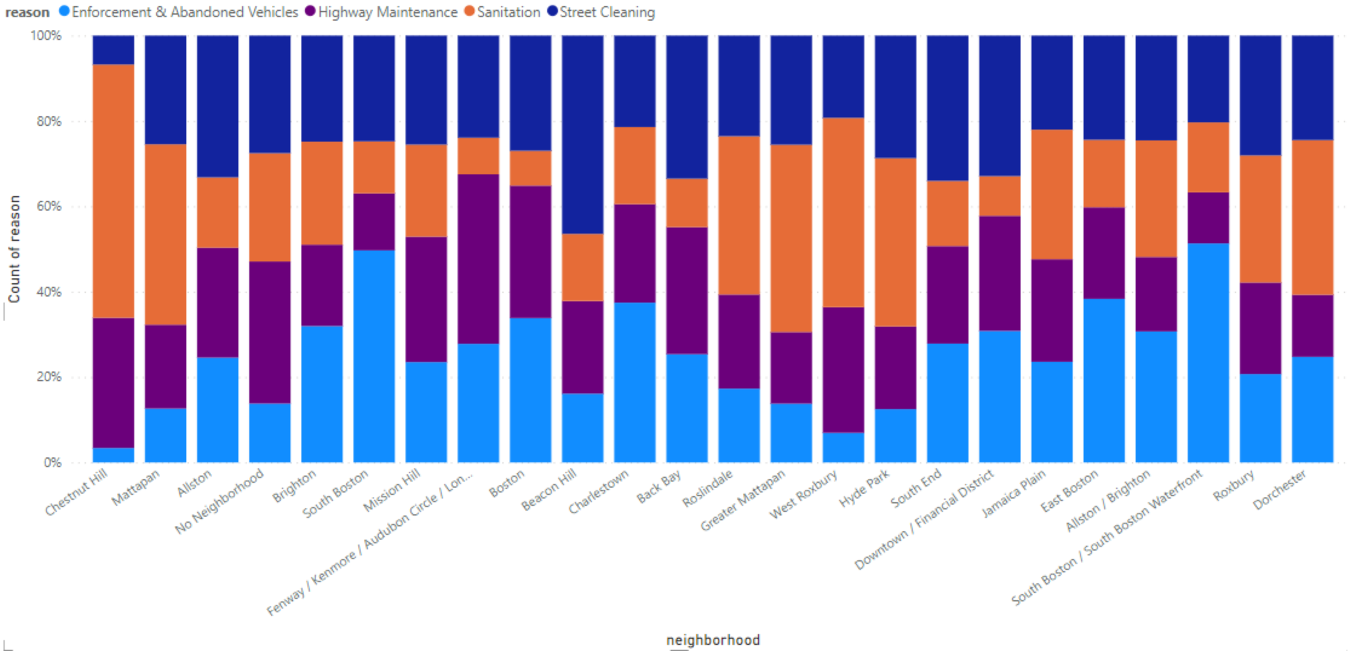






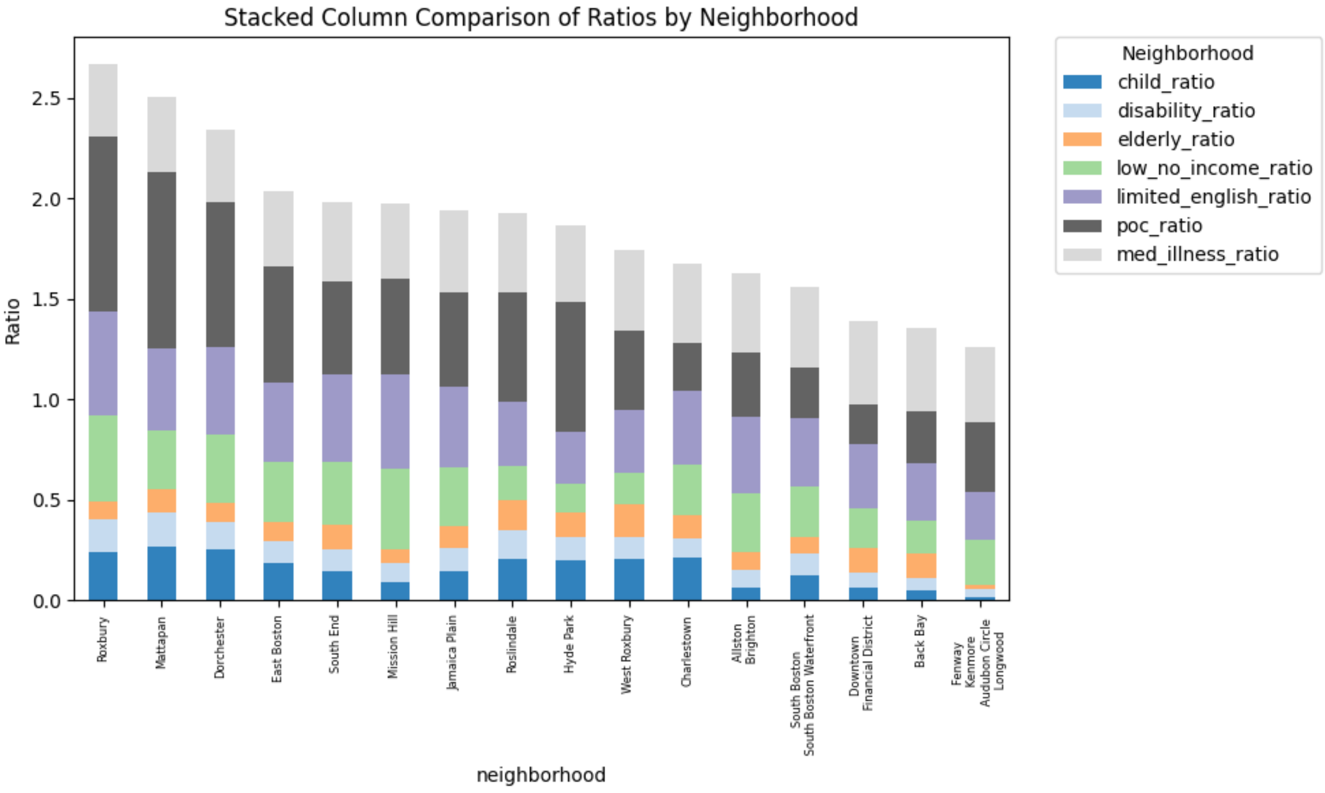
So far the only significant relationship occurs between the child ratio and median resolution time. This may support the idea that service is efficient regardless of the population. Though we need to be careful since this data is from the 2008-2012 American Community Survey 5-year Estimates (ACS) data by census tract: Black, Native American, Asian, Island, Other, Multi, Non-white Hispanics. One reason for the difference we see in the median resolution time, could be possibly attributed to the nature of the task being completed. One thing to note is that median completion time per task is not even across tasks, and that may be important in determining the median resolution time for an area depending on the composition of its request body. Note for example that "Street Cleaning" has a very low median resolution time, but "Sanitation" has a higher median time. On top of that when looking at some of the neighborhoods with higher median resolution time we can see composition differences.



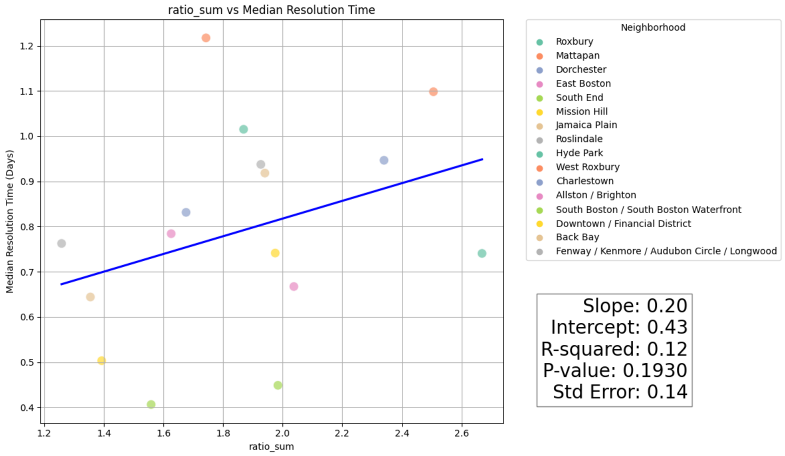


There is a possibility that those neighborhoods with a higher proportion of reasons being Sanitation tasks tend to have higher resolution times. Greater Mattapan and West Roxbury, the top 2 in terms of median resolution time, have around 25% representation by Sanitation, while Beacon Hill and South Boston, the bottom 2, have around 10%.

In order to do a first attempt at characterizing, a high social vulnerability area, we took the sum of the different ratios provided by the ACS dataset, and got the following:



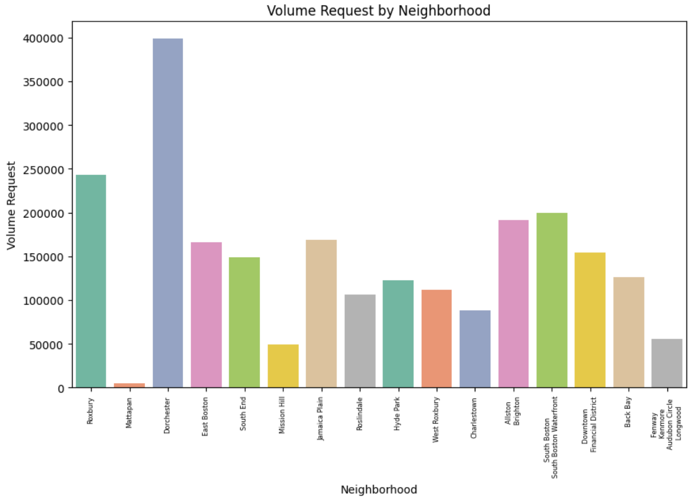
We were curious to see the relationship between median resolution time and this new feature and we got the following results:



Again, nothing conclusive in terms of a difference in efficiency in relation to social vulnerability.

Another set of interesting points to examine is the volume of requests, this was plotted while maintaining the descending order of the sum of index ratios. Unfortunately, no clear pattern emerged.

One potential avenue we were considering in terms of SVI, is to use more recent data, and a more robust scoring method for SVI. Currently we used the sum of ratios, but a more thoroughly studied index might be better suited to our analysis.



We were hoping to split our extension project efforts into two streams across our group members, and we will also explore the rate of closure for the same type of service requests, across NEIGHBORHOODS, CITY COUNCIL DISTRICTS, NEIGHBORHOOD SERVICES DISTRICT, ZIP CODE.

In terms of models, we believe the most feasible and most practical model we could build is a regression model that predicts resolution time for a task given a select number of features. This will be our primary target. Other models include predicting total future volume of requests, and potentially categorical predictions for neighborhoods/reasons/subjects/departments.

Individual Contribution

- Raul-Fikrat Azizli
  - 311 Data pre-processing and cleaning
  - Exploring and Answering Base Question
  - Interactive Base Question Notebook
  - First Version of EDA for the project
- Mahdi Khemakhem
  - Build PowerBI dataset, daily web-scraper update.
  - 311 Data pre-processing and cleaning.
  - PowerBI Report
  - EDA for SVI extension project.
- Laerk Ndreko
  - Preliminary Data Analysis of 311 Data
  - Visualization of Data
  - Exploratory Data Analysis of Extension Project
  - Presentation Preparation
- Gersian Collaku
  - Exploratory Data Analysis
  - Visualizing the Data
  - Preliminary Data Analysis of Extension Project
  - Extension project base questions
- Rohan Anand
  - Notebook base question
  - Exploratory data analysis
  - Making the notebook interactive