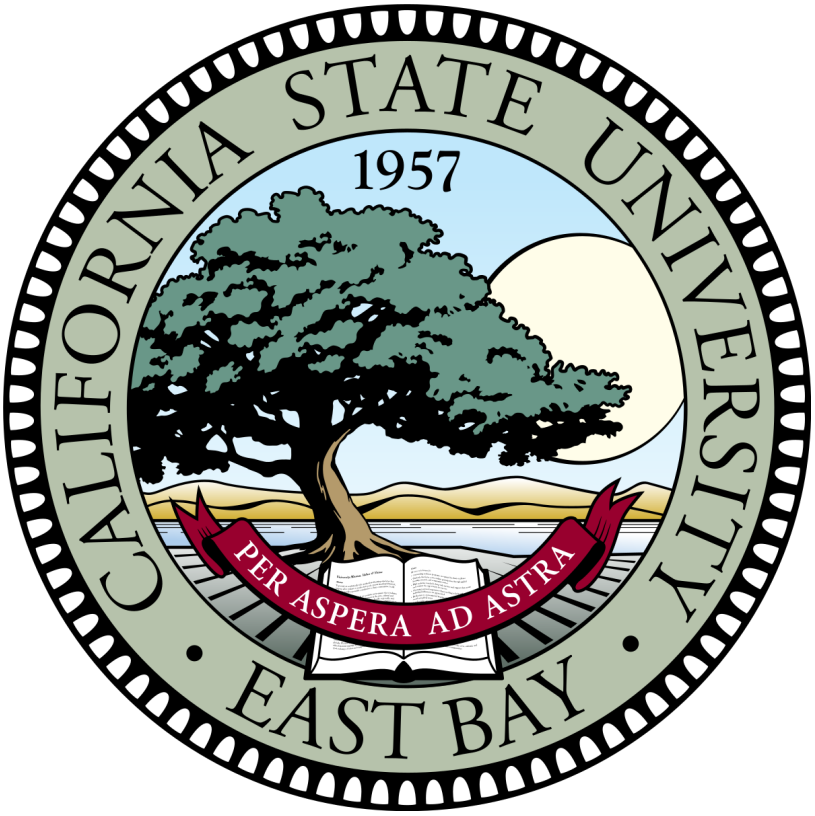


Data Visualization for San Francisco Police Department Incident Reports: 2018 - 2020

Vaishnavi Udaya Kumar
Net Id: MA5458

Supervisor: Prof. Joshua Kerr

A Mid-Term Project Report for STAT 651 (Data Visualization) submitted to the Dept. of Statistics & Biostatistics, California State University, East Bay in partial fulfilment of the requirements for the degree of

Masters in Statistics (Data Science Concentration)

Fall 2021

Table of Contents

Abstract

San Francisco Police Department Incident Reports: 2018 to 2020
Vaishnavi Udaya Kumar, California State University, East Bay

This dataset includes incident reports that have been filed as of January 1, 2018. These reports are filed by officers or self-reported by members of the public using SFPD's online reporting system. The reports are categorized into the following groups based on how the report was received and the type of incident:

- Initial Reports: the first report filed for an incident
- Coplogic Reports: incident reports filed by members of the public using SFPD's online reporting system
- Vehicle Reports: any incident reports related to stolen and/or recovered vehicles

The objective of this project is the apply the learning from the STAT 651 course to visualize the aforementioned dataset's features such that the insights of the data can be conveyed intuitively by means of charts and plots to the reader.

Problem and Motivation

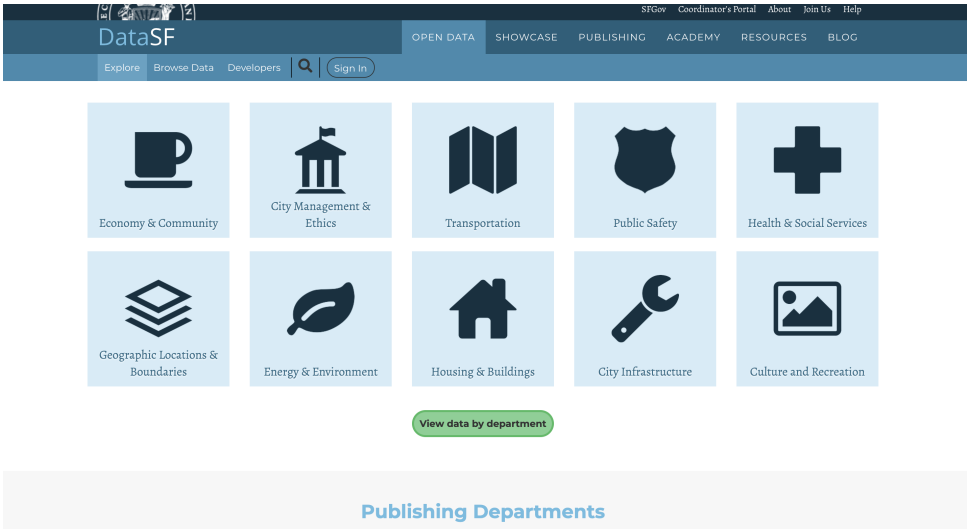
Being residents of the Bay Area, crime, especially in San Francisco has unfortunately become an infamous part of our daily news (and lives). The problem is widespread and overshadows the many good things the city is known for. Being a student of Statistics and Data Science, I'm motivated to dig into the available data about crime in the city with the hope of finding patterns, trends and correlations which can give me a better understanding of what has been happening. My goal is to convey my findings through this report by means of a visual aids which can help the reader better comprehend the state of criminal affairs in the city through the lens of data visualization.

The motivation to choose this data set, was to be able to apply the statistical concepts and techniques learnt in the STAT 651 course, to be able to analyze and data, and derive useful insights and generate meaningful data visualizations.

Data Description

0.1 Data Source

The source of the data is the *data.sfgov.org* website and contains San Francisco Police Departments (SFPD) Incident Report Datatset which is one of the most used datasets on **DataSF**. The dataset compiles data from the department's Crime Data Warehouse (CDW) to provide information on incident reports filed by the SFPD in CDW, or filed by the public with the SFPD. Data is added to open data once incident reports have been reviewed and approved by a supervising Sergeant or Lieutenant. Incident reports may be removed from the dataset if in compliance with court orders to seal records or for administrative purposes such as active internal affair investigations andor criminal investigations.



0.2 Data Features

This data set consists of 527,620 observations across 26 variables. Some of the key features we will be utilizing in this exercise are:

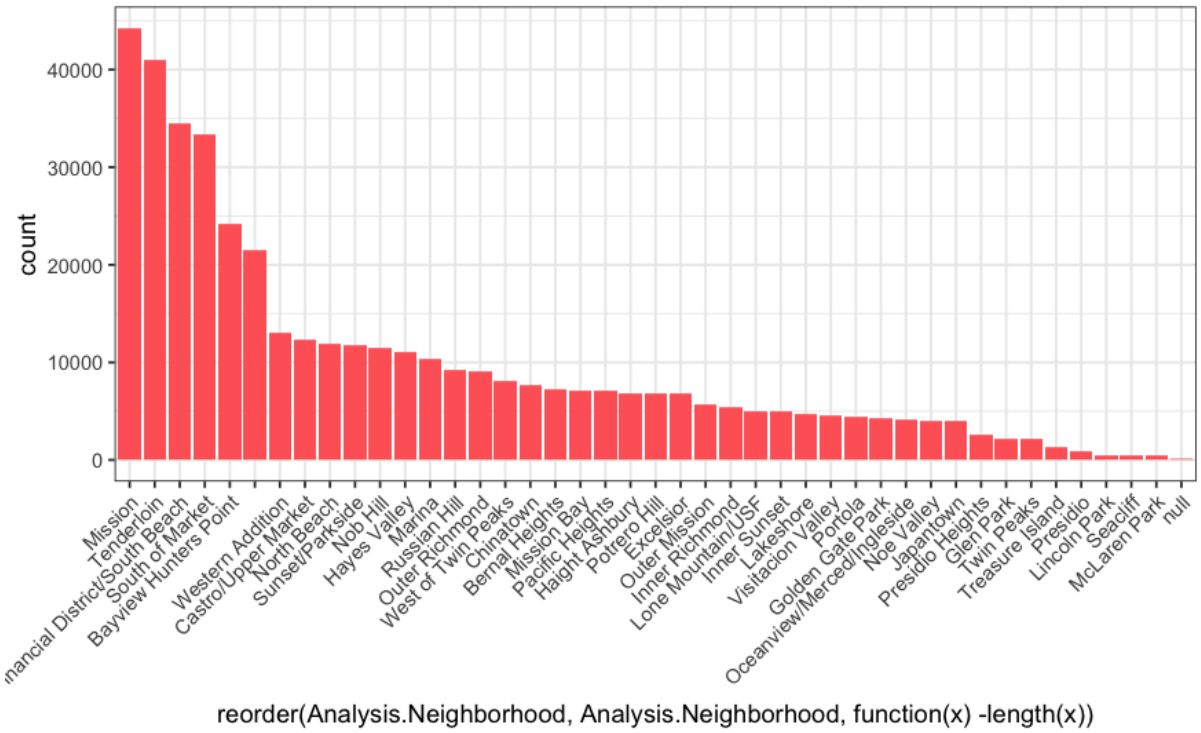
- **Incident Date**
- **Incident Category**
- **Incident Description**
- **Resolution**
- **Police District**
- **Analysis Neighbourhood**
- **Latitude**
- **Longitude**

Questions of Interest & Data Visualization

1.0.1 Question 1 - Distribution across neighborhoods

"How are the incidents of crime distributed across the various neighborhoods of San Francisco? Can we identify neighborhoods where crime rate is significantly higher and ones which are relatively safer?"

Potential Use Case: With this information, we can support stakeholders' decision making - such as potential renters or new business owners can decide to go with safer neighborhoods.

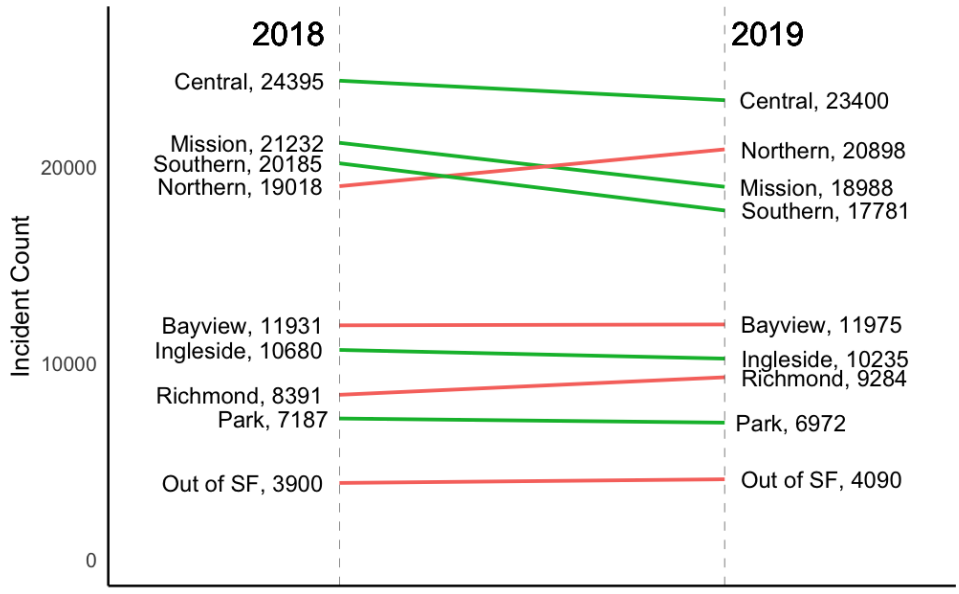


From the above bar plot, we attempt to visualize the frequency of incidents in San Francisco's neighborhoods from 2018-2020. We see that the rate of crime varies significantly from one neighborhood to the other. As seen, neighborhoods such as Mission, Tenderloin, Financial District, SOMA, Bayview Hunters Point rank high on the crime rate spectrum while neighborhoods such as McLaren Park, Seacliff, Lincoln Park and Presidio are much safer locations comparatively.

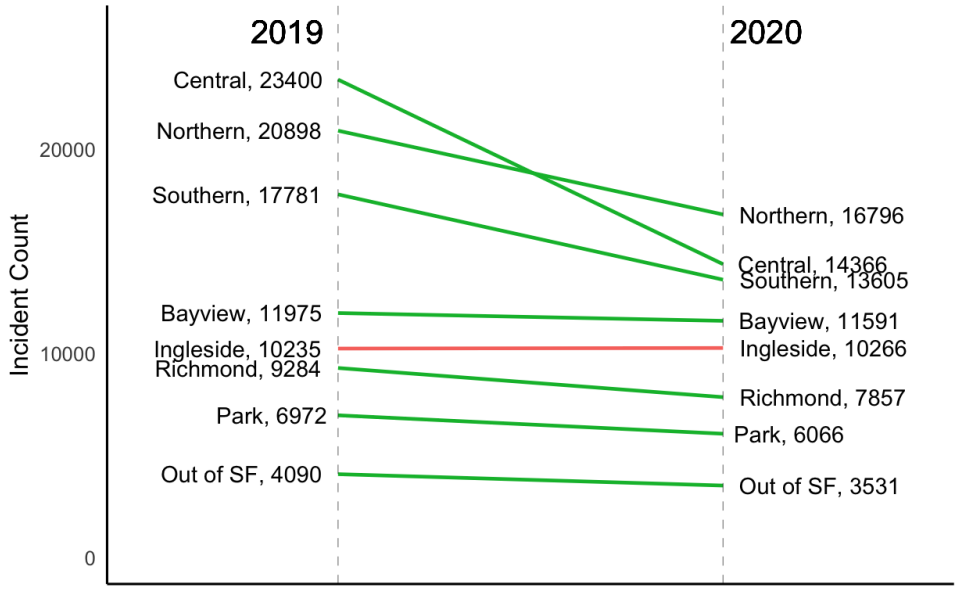
1.0.2 Question 2 - Yearly trend of neighborhoods

"How are the crime rates in the given neighborhoods changing during successive years?"

Potential Use Case: With this information, we get a clear picture of the rise, fall or stagnation of the number of reported crimes in each of the neighborhoods compared to it's numbers from the previous year. This would not only be beneficial in comparing the effect of initiatives taken by the city to reduce crime across cities, but also show the trend in which a neighborhood is moving.



The above visualization shows a comparison between the number of reported crimes in some of the active neighborhoods between the years 2018 & 2019. The trend is color coded with red representing a rise, and green representing a fall in crime rates of the respective neighborhoods. As seen, there is a slight decrease in the number of incidents in neighborhoods such as Central, Ingleside and Park. We also see a steeper decrease in crime rates in neighborhoods such as Mission. On the other hand, we see an increase in the number of incidents in neighborhoods such as Richmond Northern Station. In comparison, we also see a slight increase in crime rates reported outside San Francisco.

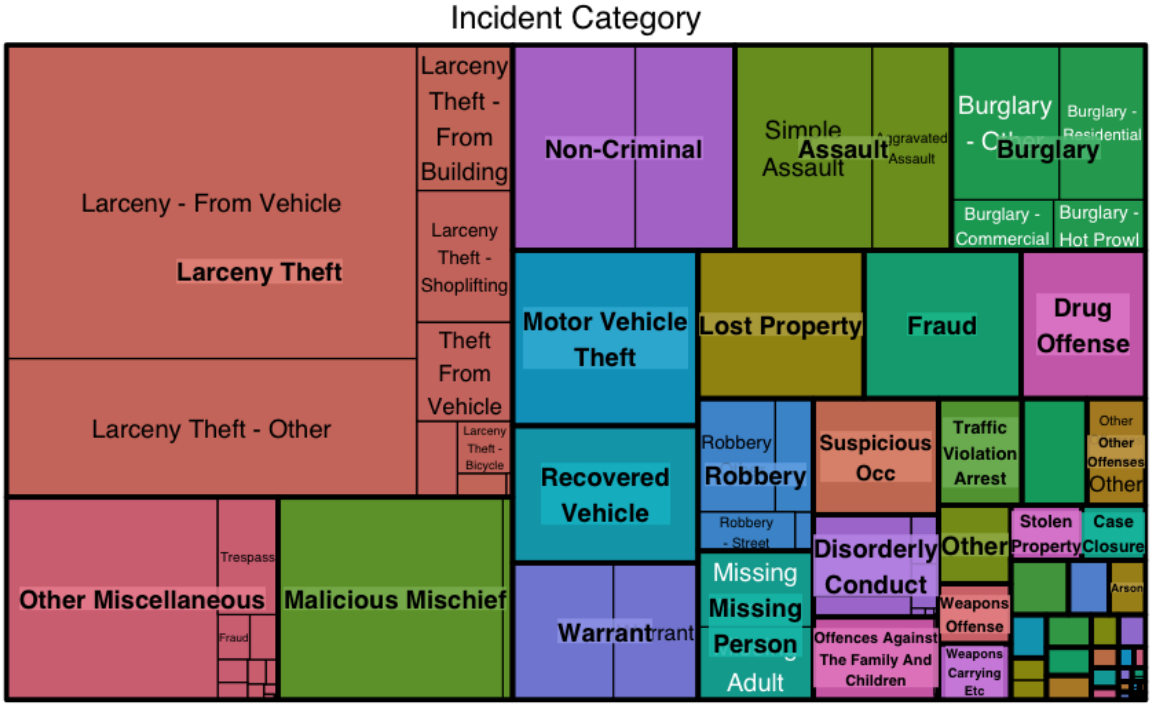


The 2019-2020 chart is very intriguing. We see a steep decrease in crime rates in almost all the neighborhoods, especially Central, Northern & Southern Station, with the sole exception of Ingleside. This could be due to the pandemic, but always good to see a drop in the crime rate. It would be interesting to see how the numbers have changed at the end of 2021!

1.0.3 Question 3 - Categories & Sub-Categories of Offences

"What are the types and sub-types of crimes being reported in San Francisco? Can we use the data to better understand the overall magnitude of such crimes?"

Potential Use Case: With this information, we can get a better idea of the relative volume of crimes in the city broken down by their category and sub-category. Over time, this visualization would be a key tool to monitor the increase or decrease of crimes in individual categories.

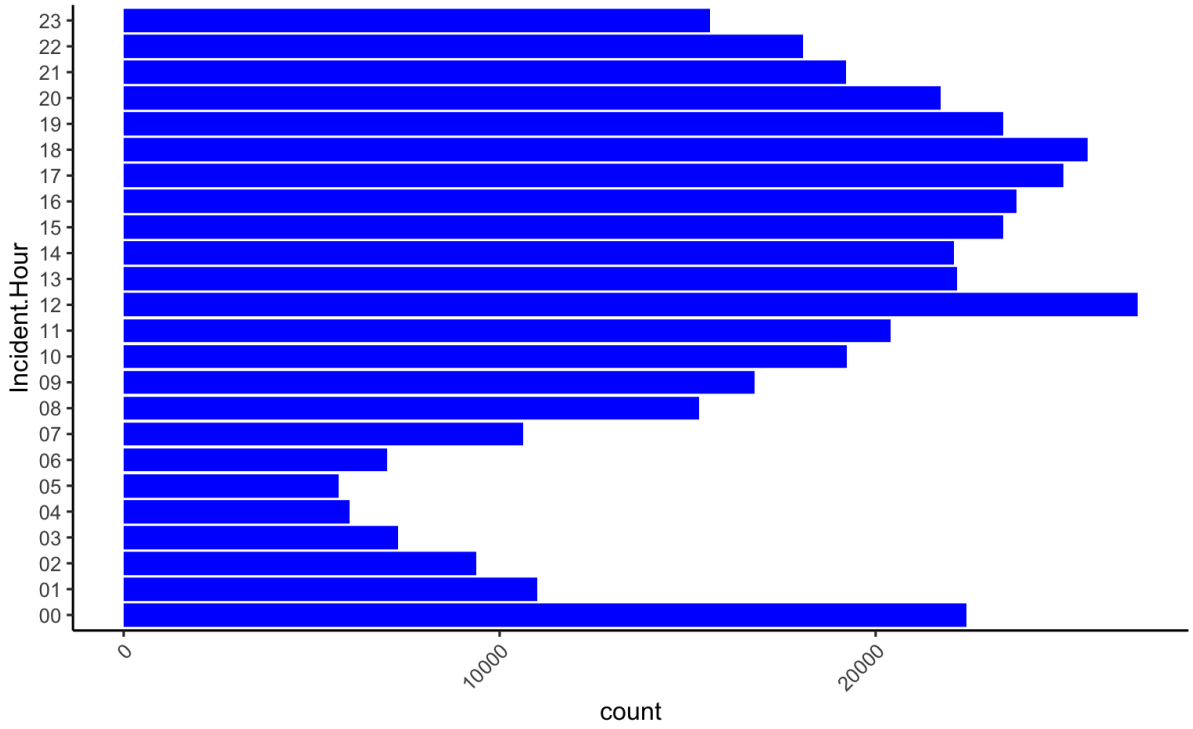


From the above tree-map, we attempt to visualize the nested relationship and volume of the categories of incidents in San Francisco's neighborhoods from 2018-2020. We see that the number of incidents in each crime category and sub-category vary significantly from one another. As seen, incidents under Larceny Theft seem to be the most common ones, within which Larceny from Vehicle seem to take up the lion's share of reported incidents. Likewise, we can see other categories such as 'Non-Criminal', 'Assualt' and 'Burglary' taking up the next few spots in this list. The advantage of a tree-map is that it is interactive, and can be zoomed into to look at more granular categories and their subcategories.

1.0.4 Question 4 - Time Component

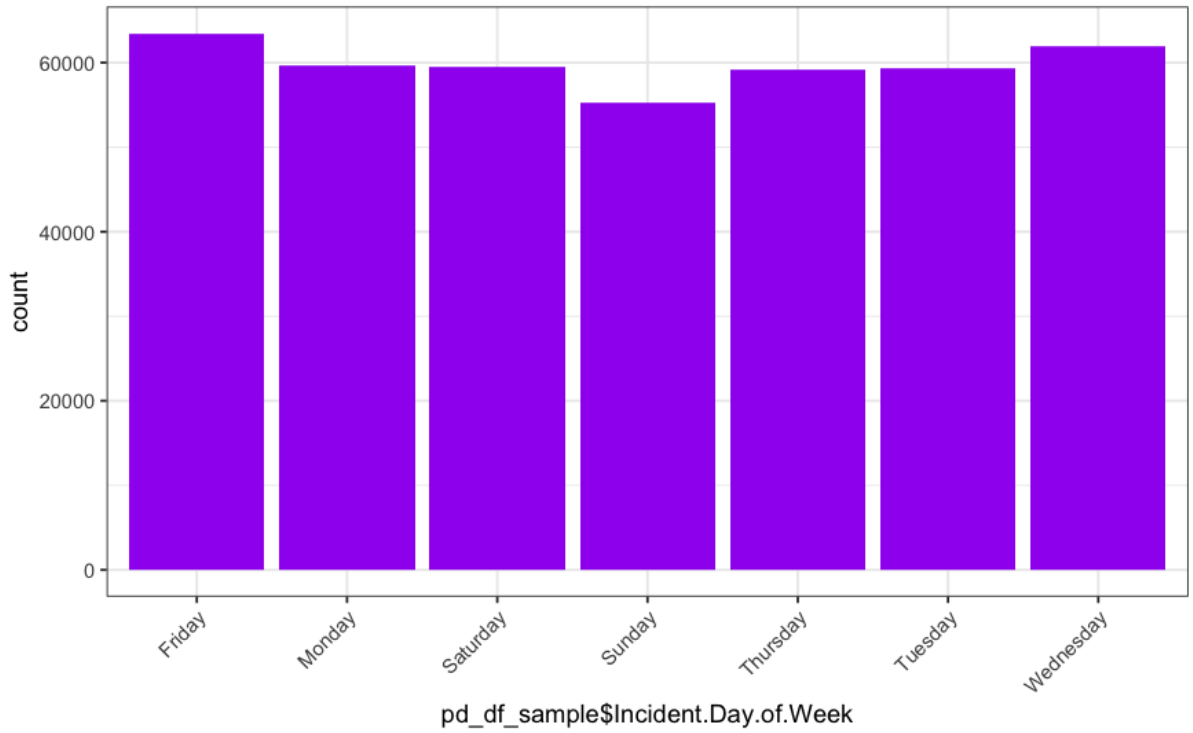
"How does time component vary with the number of reported incidents? Can we estimate the number of reports expected at a given hour or a given day? "

Potential Use Case: With this information, we can support law enforcement to be prepared to handle the increased or decreased workload efficiently.



From the graph above, we see that the number of reported incidents varies strongly with the time of the day. Here, we see a significant drop in the number of incidents during the early morning hours i.e., between 1:00 AM and 6:00 AM. Then, we see a steady increase, peaking at about midday and a second peak around 6:00 PM, before going back on the downtrend again. We do see a spike around midnight.

Let us zoom out and see how this changes with days of the week.



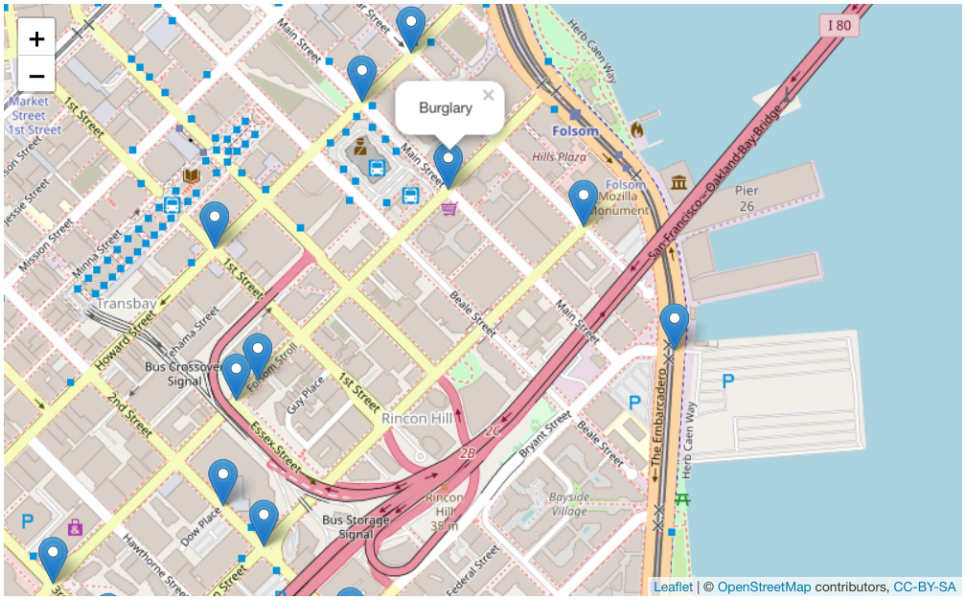
By analyzing the above graph, we see that Fridays seem to be busiest day for law enforcement, seeing the highest number of incidents, closely followed by Wednesdays. Mondays, Tuesdays, Thursdays and Saturdays seem to have an almost similar number of reports coming in. Sundays stand out as the day with the least number of incidents.

Using this information, law enforcement can plan the schedule of this personnel such that there are optimal number of officers to handle the expected load of reports at any given time.

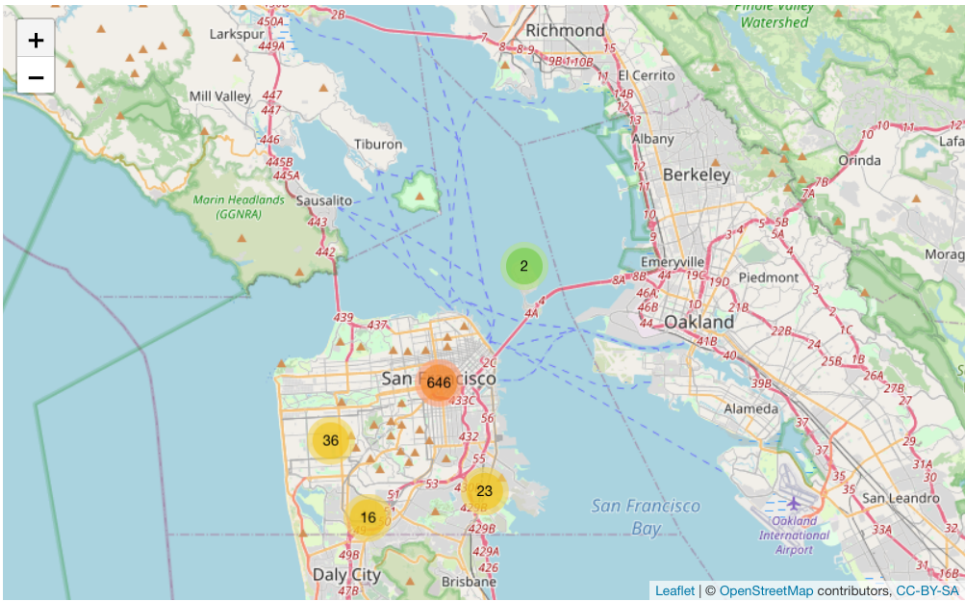
1.0.5 Question 5 - Spacial Visualization

"Can we visualize the locations where a particular crime has occurred? Can we add granularity to our visualizations and monitor the number crimes reported down to the block and street level, and color code the locations appropriately?"

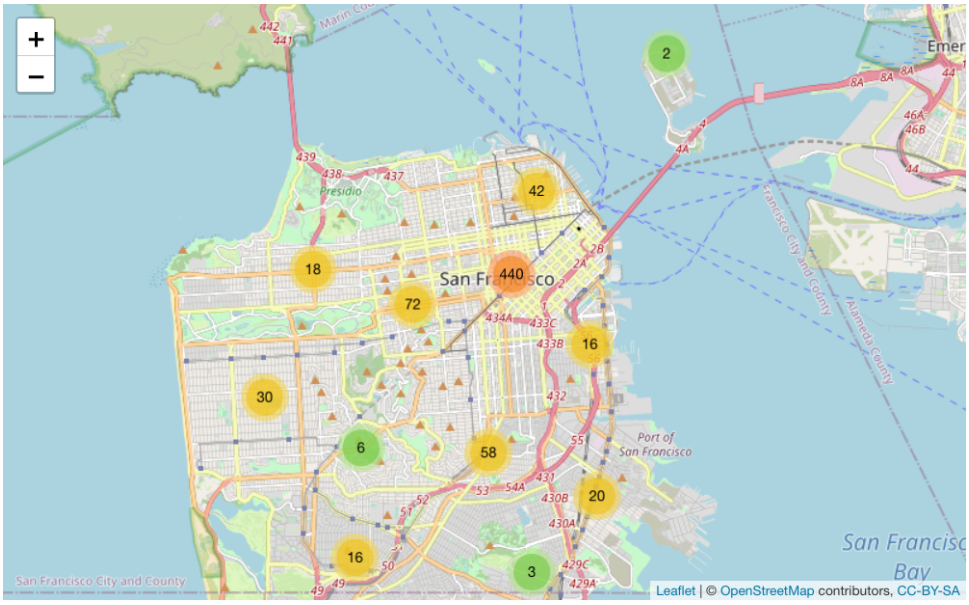
Potential Use Case: With this information, law enforcement would be able to immediately pin point locations of criminal incidents on the map accurately for deployment of personnel. This would also provide the authorities a bird's eye view of the city and help intuitively understand the rate of crime in these locations.



In the above map, we see that we can pin point exact locations where a particular type of incident has occurred, in this case, Burglary. On an interactive tool, we have the capability to change the category of offense on the fly and be able to render the map instantly with the updated locations. We can also limit the number of results based on incident date.



As seen from the second map, we can visualize the number of incidents from clusters of locations on the map. At a high level, these are clustered into fewer areas, and as we zoom in, the clusters are spread out, revealing the breakdown of the numbers coming from different locations in that area, as seen in the maps below.

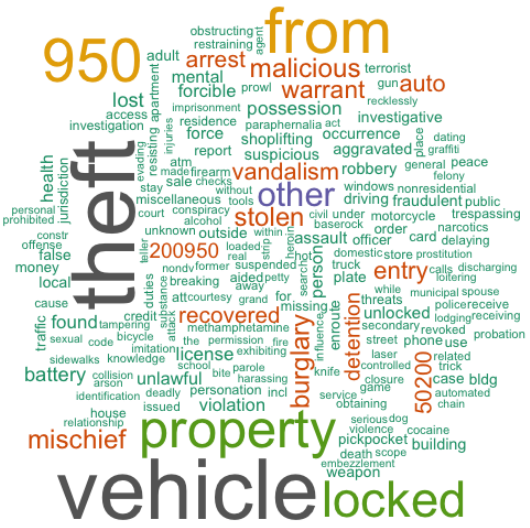


This is a further zoomed image of the same map. As seen, we can increase the granulaity of the results down to the street level and thus obtain a bird's eye view of the city. The number of incidents from each cluster is color coded such that the message of the visualization is conveyed intuitively.

1.0.6 Question 6 - Word Cloud

"From the descriptions of the reports, can we create a visualization which provides an overview of the incidents, highlighting key words which appear more often than others? "

Potential Use Case: With this information, one would be able to extract and visualize the most commonly occurring words across all the reported incidents. On a large time-frame, this could potentially bring out key words or phrases which will help identify the most pressing issues over time.



The following word cloud is generated from the text descriptions of the reported incidents contained in the three years worth of data from SFPD. We see quite a few words which stand out such as 'vehicle', 'theft', 'burglary', 'vandalism' etc. (Likewise, we also see a couple of words which might not add any value such as 'From' and 'other'. Removal of stop words from the corpus is one way to eliminate these words.) Such a wordcloud helps highlight the main areas or most common topics which are reported about, thus adding a layer of transparency to the underlying data.

References

[1] STAT-651 Lecture Slides by Prof. Joshua Kerr, 2021.
[2] *The LaTeX Companion* by Addison-Wesley, Reading, Massachusetts, 1993
[3] Data Visualization with R by Rob Kabacoff
[4] DataSF Website (data.sfgov.org)

Acknowledgement

I would like to express my heartfelt gratitude for Prof. Joshua Kerr for his continued guidance and support. I would also like to thank my fellow classmates with whom I've had the opportunity to have many constructive conversations through the course of this semester.

Appendices

Appendix 1: R Code

Link to Github Repo

(For print version of the report, please scan above QR code to view the Github code repository)