Evaluating Patient Receiving Locations

by

Clay Steward

As partial completion of the IBM Applied Data Science Capstone project on Coursera

submitted

March, 2021

I) Introduction

Background

A growing concern for many cities and communities is the potential need for additional medical treatment facilities or services. Continued increases in population, ages, and widening financial gaps are putting increasing pressure on fire department or other emergency medical service (EMS) provider agencies to fill this needed service. While most service agencies evaluate items such as call volume and response component times, they rarely feel impowered to evaluate other aspects that effect their workload. One potential issue is the availability of definitive care facilities such as doctors' offices, clinics, hospitals, emergency, and urgent care centers.

Thesis

This evaluation is an introductory evaluation of locations of facilities versus the locations of the medical incidents that an EMS provider responds. The underlying themes here are that workload for the EMS provider is increased as the distance to the medical facility increases, due to extended transport times, and location of medical facilities may have an impact on whether patients utilize these facilities directly or look for assistance from emergency service providers.

Interest

Many strategic planning executives, senior officers, and community leaders may be interested in this type of evaluation. This type of evaluation may help city or county leadership, such as fire or ambulance services leadership, or care facility owners make better decisions on the location and types of services provided. The point here is that there are more dimensions to most service analytics and there are tools available to help communities and agencies evolve.

How to deliver emergency and urgent medical care is a very important evaluation that should include multiple facets. While this paper is meant to introduce this type of evaluation to the interested service provider or leader, it must be made clear that there are too many variables at issue for this paper to be a definitive application of the processes herein. This evaluation is an academic evaluation and many of the factors related to the actual selection of urgent/emergent care providers were not readily available and so were not evaluated. Items such as available real estate, financial impact, zoning, political and social will, available staffing talent, among others are not approached or discussed. Therefore, this paper should be viewed as an introduction into using computer modeling, mathematical processes, and machine learning to begin to identify areas of weakness, not as a recommendation for action.

II) Methodology

The methodology utilized in this analysis focuses primarily on the tools and techniques introduced in the IBM Data Science Professional Certificate courses. These include data visualizations, statistical evaluations, and machine learning techniques and focus primarily on using the python computer language.

Data

The data for this evaluation is publicly available information and found in three separate data sets.

- The first set is a five-year history of emergency medical incidents where the San Francisco Fire Department responded. San Francisco maintains a publicly accessible fire department incident data set at https://datasf.org/opendata/.
- The second set is a list of urgent care and emergency room facilities from the Foursquare API data set that includes urgent care and emergency rooms. The Foursquare data set included the venue data in Foursquare. This was found by utilizing an API query to https://api.foursquare.com/v2/venues/.
- The final set is the GeoJSON file that includes all the San Francisco analysis districts. The city government of San Francisco has broken their city into analysis neighborhoods. According to their website at https://data.sfgov.org these districts are determined by capturing areas as defined by census blocks. The incident data has analysis districts already associated with it so the Foursquare data was joined using GeoPandas methodology with the analysis district data to provide a common reference point for the two data sets.

Data Joining, Shaping, and Erroneous Data Removal

This section is a brief overview of the steps taken to reshape the data into an analytical data set. Both the incident data and the Foursquare data had some level of transformation and shaping.

<u>Incident Data</u>

The incident data from the San Francisco open data portal included records for the incident and the apparatus that responded to each incident. Only medical incidents where a patient existed were finally introduced into the data frame. The first processing reduced each incident to a summary of its component parts and returned 471,962 unique incidents.

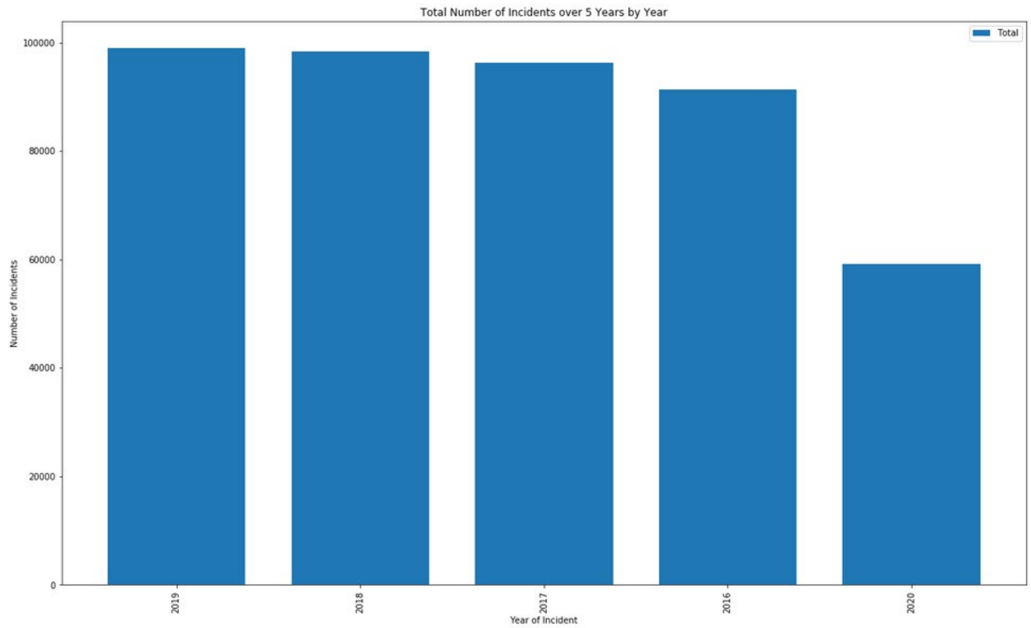

Figure 1: Count of Incidents by Year

As it turned out during the analysis there were too many incidents to provide a usable visualization. The most recent data was desirable, however the 2020 data volume appeared to be significantly less than the previous 4 years. This may be due to the Covid-19 pandemic, so this data was removed, and the data was reduced to the 2018 and 2019 data points.

Since the analysis revolved around geography, specifically latitude and longitude, anything that was outside of 36 to 38 degrees of latitude and -121 to -123 degrees longitude were discarded. The total incidents were reduced to 197,365 total incidents, 10,783 priority transports, 158,920 non-priority transports, and 27,662 non-transported incidents.

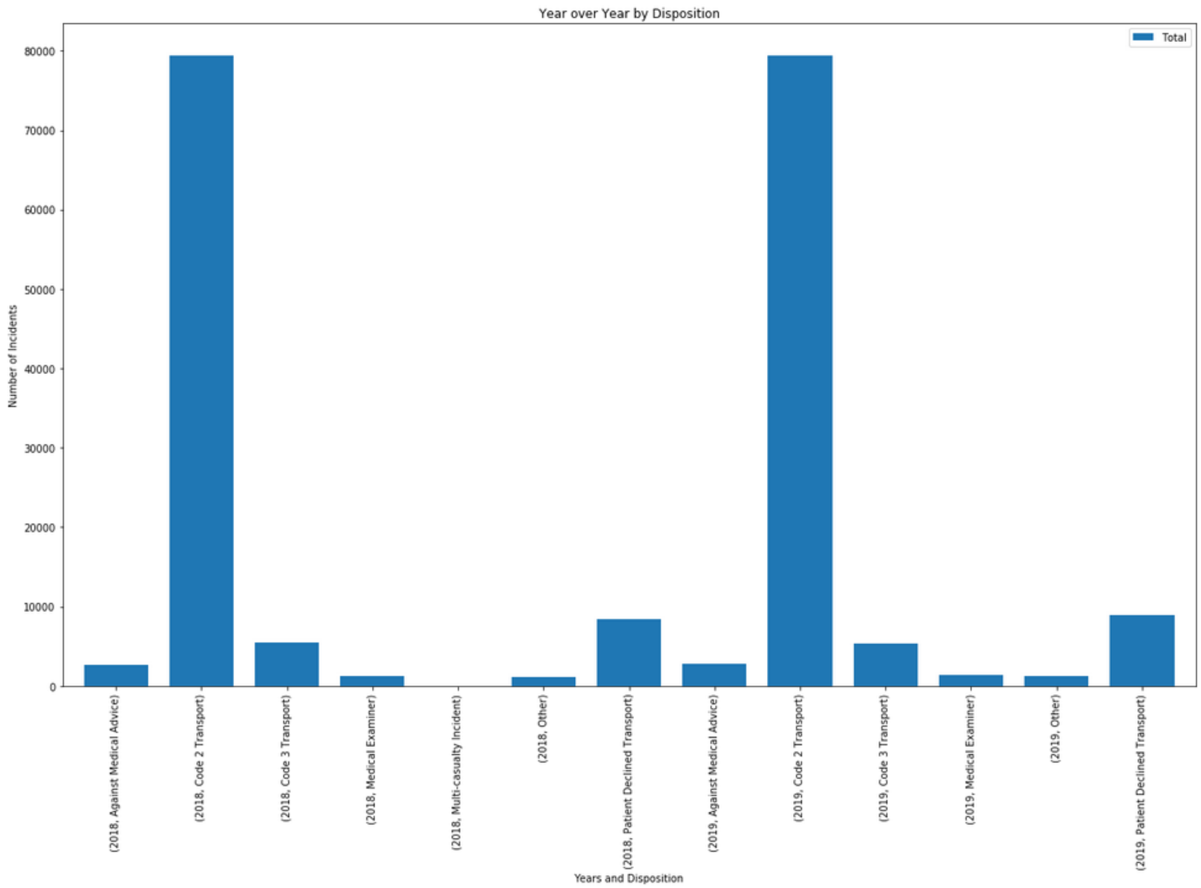

Figure 2: 2018 & 2019 Call Volume by Disposition

One additional data point was created, transport time in seconds (transporting versus at hospital). Unused column data was also reduced to create the following data frame.

	ID	Disposition	Latitude	Longitude	Neighborhood	Date	Transporting	At Hospital	Year	Transport Time
187565	18000002	Code 2 Transport	37.797611	-122.431831	Marina	2018-01-01	2018-01-01 01:11:40.992	2018-01-01 01:24:06.970	2018	745.0
187566	18000003	Code 2 Transport	37.795467	-122.396774	Financial District/South Beach	2018-01-01	2018-01-01 00:21:04.032	2018-01-01 00:21:16.042	2018	12.0
187567	18000005	Code 2 Transport	37.762745	-122.419838	Mission	2018-01-01	2018-01-01 00:26:46.003	2018-01-01 00:29:12.019	2018	146.0
187568	18000006	Code 2 Transport	37.786270	-122.412994	Tenderloin	2018-01-01	2018-01-01 00:35:14.035	2018-01-01 00:47:36.038	2018	742.0
187569	18000007	Code 2 Transport	37.771290	-122.408531	South of Market	2018-01-01	2018-01-01 00:50:01.968	2018-01-01 01:01:29.021	2018	687.0

Figure 3: Incident Data Frame Head

Four Square Location Data

Foursquare data was gathered for two specific categories and retrieved using their API. The two categories were Emergency Room and Urgent Care Facility. This query returned 82 facilities. The data frame was reshaped and formatted for clarity. When the data was joined with the analysis district data there was an unintentional but beneficial consequence. There was a change in the total facilities from 82 to 38 facilities, and by happy coincidence a few of the dropped facilities appeared to be categorized incorrectly in Foursquare, the remaining data appeared valid. While a full survey of the facilities close to the incidents should include neighboring towns and locations, for this evaluation it is more effective to only include facilities within the actual city boundaries. The final data set became 40 facilities with the following information and broken into two data sets, emergency room/hospital (17 locations) and urgent care/clinics (21 locations).

	categories	lat	lng	nhood
1	Emergency Room	37.774396	-122.422790	Hayes Valley
2	Emergency Room	37.795936	-122.400003	Financial District/South Beach
5	Emergency Room	37.768518	-122.434746	Castro/Upper Market
6	Emergency Room	37.786072	-122.423173	Western Addition
7	Emergency Room	37.747807	-122.420586	Bernal Heights

Figure 4: Four Square Location Data Head

Analysis Districts

The city government of San Francisco has broken their city into analysis neighborhoods. According to their website at https://data.sfgov.org these districts are determined by capturing areas as defined by census blocks. This will be an assist later analytics and keeps the data in buckets that the city already utilizes. This is a series of GeoJSON geographic shapes with the associated neighborhood names. The incident data already includes this information, but the Foursquare data does not. As a departure from some of the loading of the data, this will be uploaded using the geopandas library for ease of joining the Foursquare data.

Models and Tools

The data was evaluated utilizing Folium choropleth maps for both call volumes for all incidents and average transport travel times for both priority and then non-priority incidents. Since there were no transport times there is no average travel time for that data set.

Clustering of incidents grouped by priority transport, non-priority transport, and no transports were completed utilizing the DBSCAN tool. To determine the distance between data points (the epsilon in the DBSCAN model) the nearest neighbor model was utilized on each data set to evaluate the sharpest curve, or elbow, in the curve.

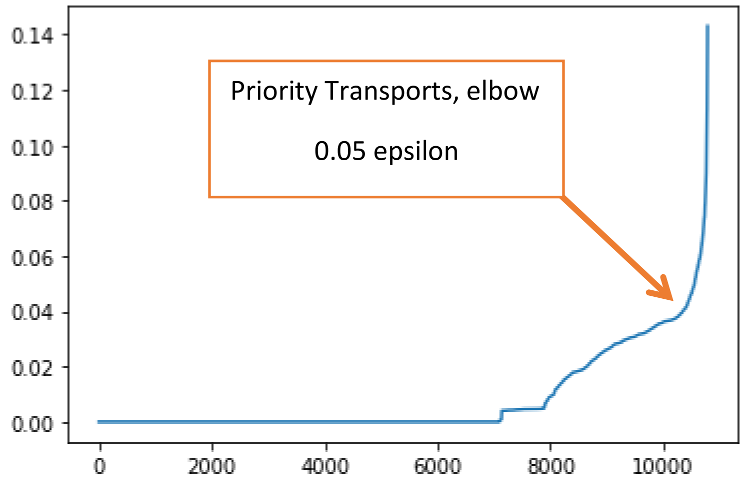

Figure 5: Example of the Nearest Neighbor for Epsilon Evaluation

The data was evaluated and sliced with three different scenarios in mind. First scenario is the priority (code 3) transports, the second non-priority (code 2) transports, and finally patient not transported.

III) Results

This data evaluation is again an introductory use of data science techniques for emergency services. The primary focus was on how the emergency medical service (EMS) is best able to cope with the demands of the customers and patients. The data evaluated shows a few areas of potential concern for those customers based on the ability of the EMS provider, the San Francisco Fire Department, to transport patients in a timely manner.

While there are certainly more dimensions that should be evaluated than the two primary dimensions evaluated here, location of receiving facilities and call volume, it appears that there are some areas of possible improvement for the service.

Call Density

The first parameter to evaluate is the call density. The San Francisco system is very busy, primarily in the city centers. The lowest call volumes are in the park areas and on Treasure Island. The highest density is in the old quarter (Tenderloin neighborhood), downtown areas, the south east Bayview neighborhood, and the western Sunset/Parkside neighborhoods. It is interesting to note that there are few resources on the west side and south side of the city, while both areas do have a significant number of incidents.

The following map shows an overall call volume by analysis neighborhood. All facilities where patients can be transported are identified as a blue dot.

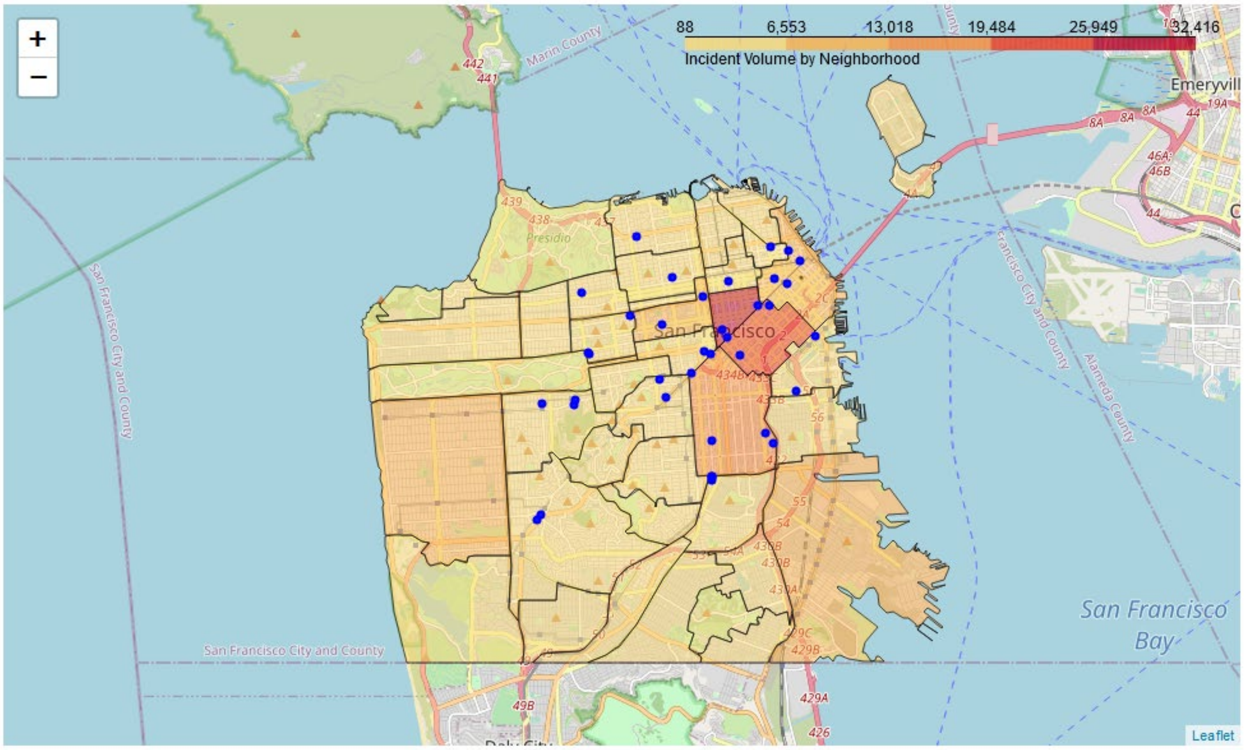

Figure 6: Heat Map by Analysis Districts of Call Volume

Transport Time as an Analysis Parameter

In emergency services one of the primary functional analysis for performance is time. For instance, how long a certain action takes or how long for a response are often used to evaluate system performance. It is outside of the scope of this analysis to go into the reasons for this key performance indicator, however it is an important consideration when looking at medical facilities and whether they may be placed in the correct location. This evaluation uses travel time as the indicator for facility accessibility.

The facilities are be broken into two sets. The first set is the emergency room/hospital set and the second set will include urgent care, doctor facilities, and other medical facilities. The reasoning for this is the assumption that an emergency room/hospital will be able to provide a higher level of emergency care than the other care facilities.

The transport time variable be evaluated in two categories. Code 2 (non-priority) transports and code 3 (priority) transports. The assumption is that the more acute or sever cases will be transported quickly and transported to the nearest accepting emergency room or hospital. The non-priority transports could potentially be transported to any facility, depending on policy. It bears repeating here that it is unknown from the data whether the city of San Francisco allows transports to other facilities, but for the sake of this evaluation the assumption is that they do allow this type of transport. The final group, the patients that were not transported will make up a different evaluation at the end of this analysis.

Code 3 Transport Averages with Emergency Facilities

The left side map shows transport time of priority transports and includes all 5 years of the data. As noted above this is the primary mode of transport within the San Francisco fire service. Also as noted in the data prep area the excessive or missing transport times have been removed. The longest transport times (red on the following charts) are between 14 and 16 minutes (848-952 seconds) on average and these areas are in the northeast and southeast analysis districts. The second chart is a DBSCAN clustering analysis of the same incidents.

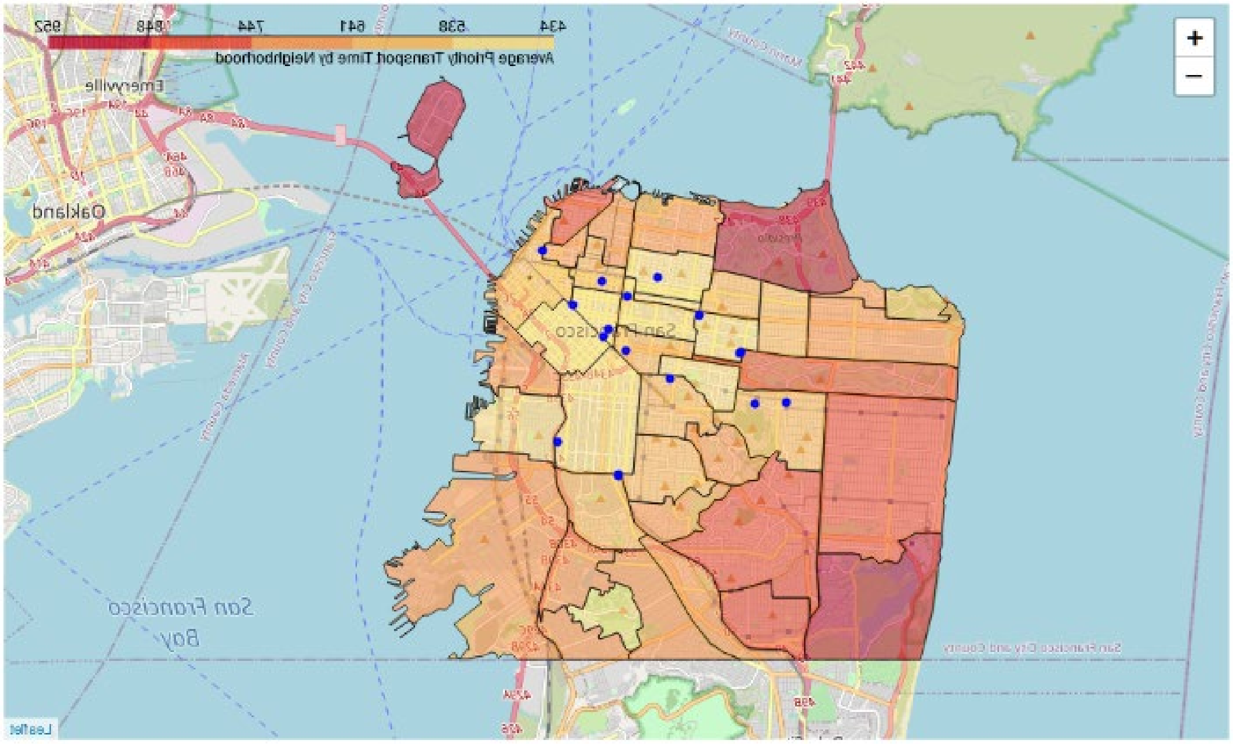

Figure 7: Average Priority Transport Times

The second chart is a DBSCAN clustering analysis of the same incidents. It is interesting to note that while many of the incidents are clustered around the emergency facilities, the blue dots, the areas that are clustered in the south and west districts have no closely tied facility, while those areas with faster transport times are central areas with access to many facilities and tight clusters.

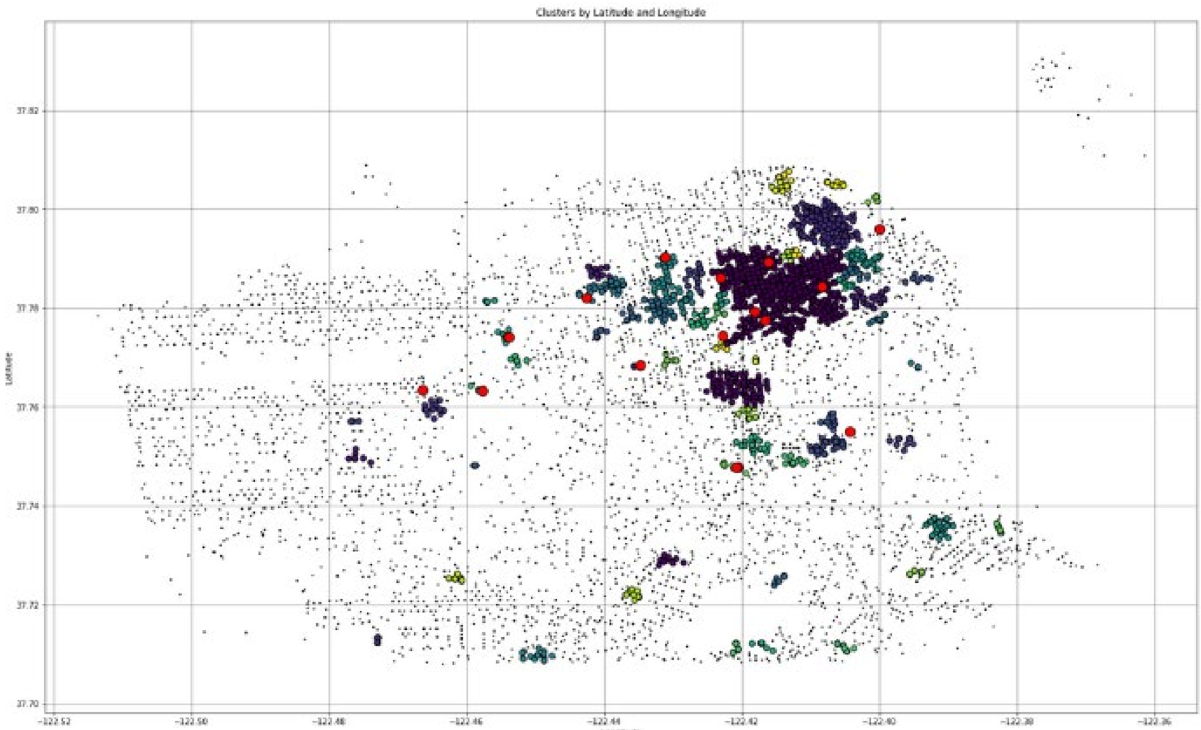

Figure 8: DBSCAN Clusters (59) of Priority Transports with ER Facilities

Code 2 Transport Average Time with All Facilities

The map shows transport time of non-priority transports. As noted above this is the primary mode of transport within the San Francisco fire service. Also as noted in the data prep area the excessive or missing transport times have been removed. The red dots are the emergency facilities while the blue dots are the urgent care facilities. The longest transport times (red on the following charts) are between 20 and 26 minutes (1234-1574 seconds) on average and these areas are in the northeast and south analysis districts.

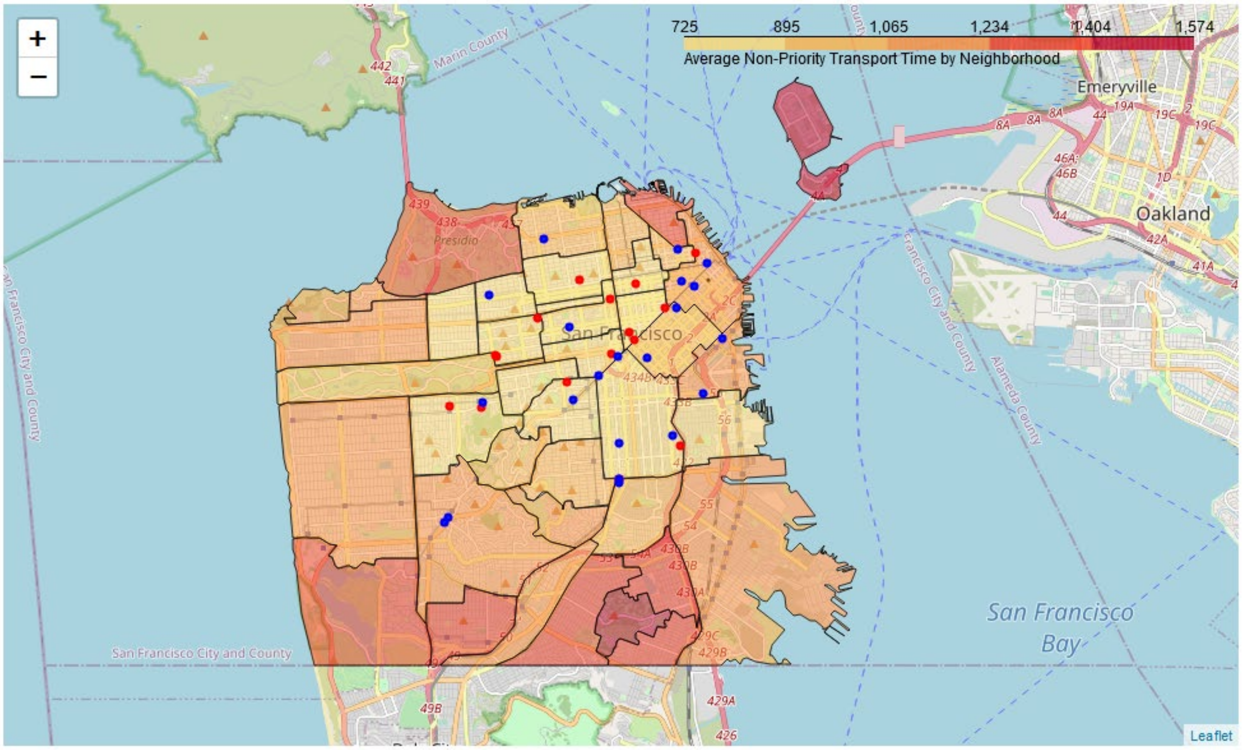

Figure 9: Average Non-Priority Transport Times

The second chart is the results of the DBSCAN clustering analysis for these non-priority calls. Again, the same phenomenon presents itself with clusters in the south and north east not near facilities, with an additional set of clusters in the west of the city that also have no apparent facilities.

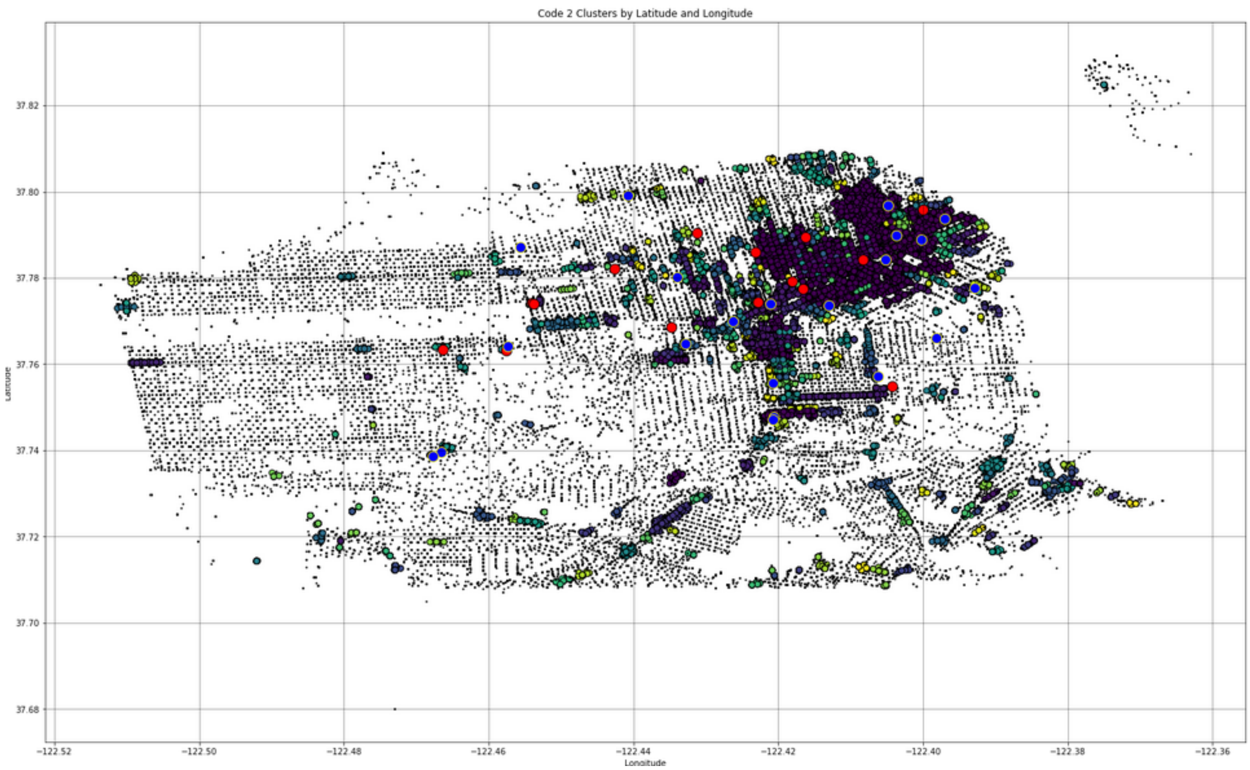

Figure 10: Figure 8: DBSCAN Clusters (370) of Non-Priority Transports with ER/UC Facilities

Non-Transport Incidents

One final DBSCAN model was created using the incidents that did not transport. Since the transport time cannot be evaluated the location and clustering of these incidents show similar underserved areas as the non-priority transports. These remain the south, north east, and western portions of the city.

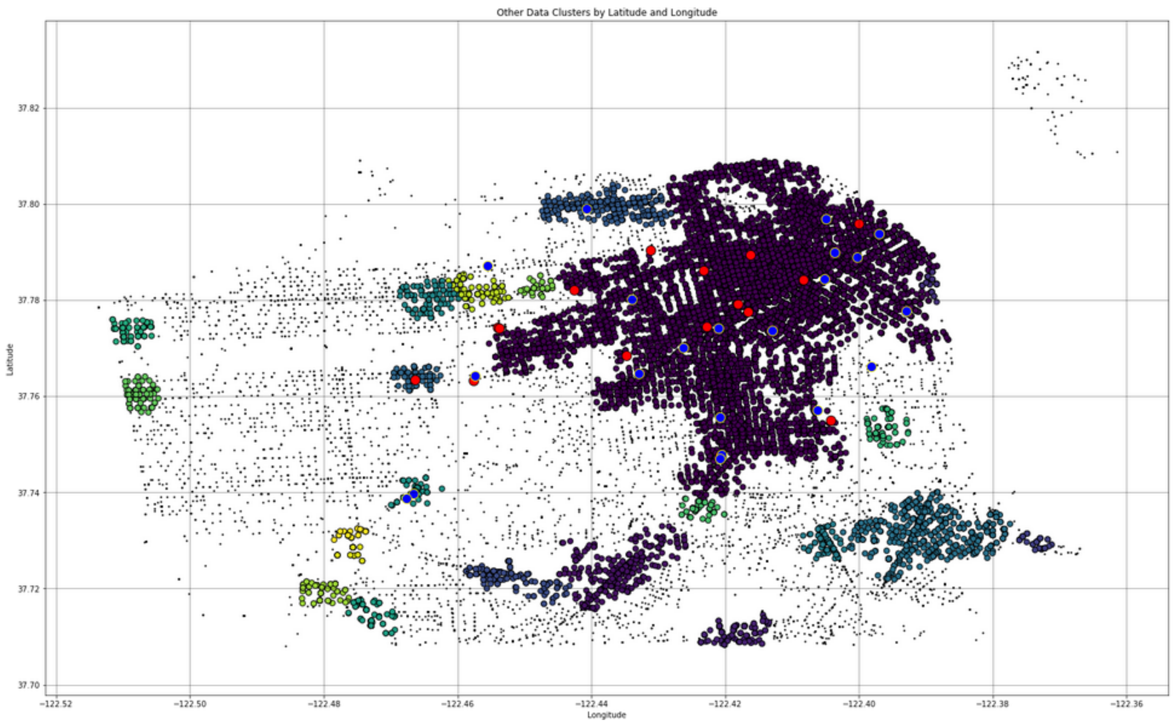

Figure 11: DBSCAN Clusters (22) of Non-Transported Individuals with ER/UC Facilities

IV) Discussion

It is difficult to make any firm conclusions based on so few parameters, however it does appear that extended transport times may have a correlation to the location of the receiving facilities. It is clear the western and southern analysis districts have no facilities close, significant clustering and call volumes, and have more extended transport times. Some of the park areas have extended transport averages but have very few data points and no clustering. It seems likely that one area of performance enhancement may be the evaluation of locations more to the south and west of the current centers. However, as noted earlier, facilities in other cities may be available for transport, so the southern analysis districts may indeed have facilities close in neighboring communities, that would not show up with this evaluation.

An anomaly is the north eastern neighborhood, the North Beach area, that does have a facility close but has extended transport times. This might be indicative of a lack of capacity of nearby facilities to handle call volume which would require ambulances to transport to other, further, locations. It also might be due to access into the area or any number of other issues. However, that analysis is outside the scope of this article.

An additional consideration when looking at this data and this analysis is the data itself. There is high confidence in the information about the incidents, the incident geography, the shape of the districts, and other data from the San Francisco open data portal. The Foursquare data is likely to be complete, however there may be some missing facilities, but the data was not audited against any official city information.

One additional talking point is the assumptions that were made. While all of them appear reasonable, they are by no means definitive. For instance, where the service can transport, what the community expectation for transport time to the hospital is, and many other variables may not have been addressed here, and this was intentional.

V) Conclusion

The reader should keep in mind that this analysis was intended as an introductory study of utilizing current data science tools to evaluate emergency services, specifically the emergency medical services, and their ability to provide timely and effective transports to their clients. While it may appear from this presentation that there are some areas of improvement available to San Francisco, and it appears that it may be due to the position and type of receiving facilities for patients, this is not a recommendation nor an accusation of any service shortages. In short this is not a conclusive suggestion, rather a different view of the information and a purely academic application of publicly accessible data.