Correlating COVID-19 Cases with Neighborhood Venues in San Francisco

IBM Applied Data Science Capstone

Introduction

The city of San Francisco has been one of the earliest responders to the COVID-19 pandemic in the United States, issuing stay-at-home orders to the people on March 14, 2020. The state of California would follow to issue a state-wide call to stay-at-home on March 20, 2020. Due to early response and strict guidelines, San Francisco is one of the large metropolitan areas in the US that has been keeping COVID-19 largely under control, with a relatively low number of cases and deaths compared to its population (7,000 cases and 64 deaths out of over 800,000 residents).

This project aims to understand the relationship between confirmed COVID-19 cases and San Francisco neighborhoods. As the city continues to re-open in the recent months, it is imperative to understand the relationship between the number of confirmed COVID-19 cases and neighborhood composition, particularly its venues. Under the assumption that most individuals are infected outside of their home, we can consider each venue as a potential site of infection. Doing so, we can analyze the relationship between the types and numbers of venues in a neighborhood and its cases.

The results of this analysis would be invaluable for local policymakers looking to understand the impact of re-opened venues on COVID-19 cases. This will inform them in shaping re-opening policy for the city in order to maintain public safety while still stimulating the local economy.

Data Source

In order to correlate San Francisco COVID-19 cases and venues, we will be using two data sources: Four Square and DataSF.

Four Square is a location technology platform that provides information on venues. It uses crowdsourced data to provide information on venues around a point of interest. The venue information consists of:

- Venue name
- Venue address
- Venue type
- Venue tips
- Venue photos

DataSF provides public datasets to the city departments of San Francisco. The dataset we will be using, will detail:

- Medical provider confirmed COVID-19 cases

- Medical provider confirmed COVID-19 related deaths
- Neighborhood population

Methodology

In order to perform the following analysis, we will first need to visualize and display the data in order to get a sense of what is happening in San Francisco. To do so, first we will be visualizing the COVID-19 cases in each neighborhood in San Francisco using a heat map. This will tell us where are hotspots within the city.

From here, we will then look at the venue data provided by Four Square and examine the top venues in each neighborhood. This will give us a sense of what is popular and where people would congregate if they were to go out in these neighborhoods. These top venues would serve as the most probable site of infection should one occur in San Francisco.

Pulling San Francisco COVID-19 data

For this project, we will be using the "COVID-19 Cases and Deaths Summarized by Geography" dataset which is provided by the Department of Public Health of San Francisco. DataSF provides an API link for users to directly download this datasets. The data is segregated based on zip code, neighborhood and census districts. For the purpose of this analysis, we will focus on the subset detailing neighborhood cases.

Identifying SF Neighborhoods

From the COVID-19 data, we obtain a list of San Francisco neighborhoods. The next step would be to find the coordinates of each neighborhood. To do so, we will leverage the Google Maps Geocoder API to search for the neighborhoods' latitude and longitude. This will provide us with a central point for the neighborhood which we can then center the venue analysis around.

Visualizing SF COVID-19 Cases

Using Folium, we can visualize the COVID-19 cases by neighborhood to see which neighborhoods have the highest number of cases. From the heatmap below, we see that the Mission and Bayview Hunters Point has the highest number of COVID-19 thus far.

Understanding Neighborhood Venues

Now that we have an idea of the number of COVID-19 cases in San Francisco neighborhoods, let us use the Four Square API to understand the kinds of venues that exists within these neighborhoods. Here, we will call the API to find the top 100 venues near each neighborhood, encode them using one-hot encoding and then list out the top 10 most popular venues for each neighborhood.

Analysis

Now that we have visualized the number of confirmed COVID-19 cases and the venues for each neighborhood, we can see if there is any correlation between them. The simplest way to do this is to segment the neighborhood into clusters. Clustering will group the neighborhoods with similar neighborhoods based on the dataset of interest. In this case, we are interested in seeing which neighborhoods are similar based on thir local venues.

From here, we can take a look to see if there is any correlation between the clusters we identified and COVID cases.

Segmenting Neighborhoods based on Venues

Now that we have identified the number of confirmed COVID-19 cases and the venues in each neighborhood, let us cluster the neighborhoods to see which are more similar to each other. From this, we can see if the concentration of COVID-19 cases is related to the venues of a particular neighborhood. To do so, we will cluster the neighborhoods based on the venue information using k-means.

Comparing COVID-19 Cases by Clusters

Now that we have clustered the neighborhoods, we can compare the COVID-19 cases per clusters. First, let us visualize the number of cases in each cluster. To do so, we will generate a box-and-whisker plot to see the spread of COVID-19 cases by cluster. From the graph below, it looks like there is no relation between the clusters and the number of cases since they vary quite largely within the groups.

Discussion

From our analysis, we see that across San Francisco, the majority of reported COVID-19 cases occur in the Mission, Bayshore Hunters Point, Excelsior and Tenderloin. These areas tend to be the more populated scenes in San Francisco where people tend to gather socially. The Mission is a well-known spot for bars, clubs, and Dolores Park. Bayshore has a good number of essential businesses and Excelsior is the location of City College of San Francisco and McLaren park. Meanwhile, the Tenderloin is a poorer neighborhood with a large homeless population which makes it susceptible to the spread of COVID-19.

When we clustered the neighborhoods based on the venues present, we get 5 clusters based on neighborhood similarity. However, the majority of the neighborhoods are withint 3 clusters while the other 2 clusters are sparse and could potentially account for outliers.

A limitation of this analysis is that it does not take into account the movement of people. The Bay Area and San Francisco has a phenomenon known as super commuters - individuals who travel a great distance to get to their workplace. This is commonly seen in the lower-income population which cannot afford to live in San Francisco but work their due to job availability or higher incomes. The inverse is also seen as many SF residents work for large technology companies around the Bay (Google in Mountain View, Facebook in Menlo Park, Apple in Cupertino, etc.). In this case, they travel and spend most of their days away from their SF homes. This analysis fails to take into account any movement that SF residents may take as a part of their job, which could lead to an infection occuring elsewhere but recorded for a SF neighborhood.

This analysis is also limited because it does not take into account the gradual re-opening and the state of the venues in each neighborhood. That is to say that it does not factor in when venues re-open. The assumption is that at this time period, venues have opened and are now a source of infection. However, it does not take into account when the venue has opened, at what capacity and how long it may have had a chance to be a site of infection.

Conclusion

The purpose of this project is to understand the local spread of COVID-19 in San Francisco under the hypothesis that open venues under the city's re-opening plan are a local site of infection. If this were true, we could understand what businesses and venues pose high risks of infection to their patrons and identify the venue make-up that makes a neighborhood most suscpetible to a spike in infections. From this information, policymakers and decision makers can carefully craft guidelines to inform businesses how they should approach re-opening and what their risks are. This would also inform how the city should prioritize businesses as they re-open in order to maintain public health and safety.

From the data, we see that their is no visible correlation between a neighborhood's venue make-up and their number of confirmed COVID-19 cases. The data shows that the different clusters of similar neighborhoods have a broad range of COVID-19 cases. Given the results and limitations listed above, we segregate neighborhoods

based on their risks of COVID-19 based on the venues available. Further testing and analysis are needed to extrapolate a conclusion from this data.