What is the relative risk of contracting COVID-19 at venues within different neighborhoods in San Francisco, CA?

Benjamin Cruz

November 1, 2020

Introduction

I. The Business Problem

As U.S. citizens adapt to the new normal of social distancing and mask-wearing, visiting new venues presents unique risks. For example, you want to go out to eat at a new restaurant but are concerned about the risk of contracting COVID-19 in the restaurant's neighborhood. While the restaurant itself may have adequate safety protocols in place, there is still the risk of contracting COVID-19 while walking or traveling through the restaurant's neighborhood.

So that is the problem: when utilizing location-based apps such as Google Maps, Lyft, or Uber, the app user does not know the relative risk of contracting COVID-19 in the restaurant's neighborhood they wish to eat at. we aim to solve. Knowing the relative risk of contracting COVID-19 while traveling to a venue would permit users to make safer decisions when deciding where and how they should eat restaurant food (such as ordering food by delivery or picking a different restaurant).

II. The Solution

To solve the problem mentioned previously, we need to assign each restaurant's neighborhood a relative risk of contracting COVID-19 score. We utilized k-means clustering to group neighborhoods by their venue type and relative risk of contracting COVID-19 scores. For example, users can see the relative risk of contracting COVID-19 scores if they want to visit neighborhoods with many Grocery scores. We can solve this problem using Foursquare's API and information from OpenDataSoft, the City of San Francisco, and the San Francisco Department of Public Health.

III. The Business Interest

The target audience and who would care about this problem are San Francisco residents, who would now be more informed on their relative risk of contracting COVID-19 when visiting the neighborhood of a grocery store or a new Mexican Restaurant. This additional functionality would also help consumers decide on how to support their local businesses. For example, should they eat in person at the restaurant or order delivery? Should they go shopping in person, or should they have their groceries delivered?

Data: Sources & Wrangling

I. Data Sources

This project utilizes four data sources: **1.) Foursquare Venue Data**, **2.) Opendatasoft.com data for the state of California, 3.) City of San Francisco COVID-19 Cases and Death Summarized by Geography data,** and **4.) San Francisco Department of Public Health.**

1.) Foursquare Venue Data:

One can use Foursquare's API to access exhaustive amounts of venue data around the world. This data ranges from the name of the venue's neighborhoods, their latitude and longitude of the venue, to the number of reviews for each venue. We accessed the San Francisco venue category, neighborhood names, and location data from Foursquare's API for this project.

2.) OpenDataSoft Data for California:

OpenDataSoft is a French company representing the data sharing platform that teams in various industries utilize to access, reuse, and distribute data that grows businesses. The company provides free data for various worldwide metrics, from bike stations data to Airbnb listings. This project took U.S. Zip Code Latitude and Longitude data for California from OpenData soft (the link: https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/table/?refine.state=CA).

3.) **City of San Francisco COVID-19 Cases by Geography and Date:**

The City of San Francisco, CA, provides publicly available structured data for various metrics regarding the city. This includes data that ranges from the energy performance of existing buildings to police department incident reports. We utilized the San Francisco COVID-19 by Geography and Date Data with our analysis. This particular dataset was from 03/03/2020 to 10/25/2020 and included zip codes and the number of new COVID-19 cases per day (rate) data (the link: https://data.sfgov.org/COVID-19/COVID-19-Cases-by-Geography-and-Date/d2ef-idww). Using this dataset in practice, one would continually pull the new COVID-19 data from the previous day to perform analysis for the current day.

4.) **San Francisco Department of Public Health, Burden of Disease and Injury Study:**

The San Francisco Department of Health conducted a study on the burden of disease & injury in the city. They wanted to see if the health of the San Francisco neighborhoods differed from one another. We took the neighborhood names and associated zip codes from one of their datasets (the link: http://www.healthysf.org/bdi/outcomes/zipmap.htm).

II. Data Wrangling

Given four datasets, we performed data-wrangling (or cleaning) for each of the four data sources.

San Francisco Department of Public Health, Burden of Disease and Injury Study Data Wrangling:

We began wrangling the San Francisco Department of Public Health data using the Beautiful Soup python package to extract the HTML table within the web page to a dataframe (full code link here). We removed the first and last row of the extracted pandas dataframe since they did not contain valid zip-code/neighborhood combinations. This gave us our initial zip-code/neighborhood dataframe that we would use later when wrangling our data.

OpenDataSoft Data for California:

We now needed to attach latitude and longitude data to the zip-code/neighborhood dataframe. We did this by exporting the U.S. Zip Code Latitude and Longitude for California (link here: https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/table/?refine.state=CA) as an excel (.xls) file. We then read the contents of this file into a dataframe. We dropped the "State," "Timezone," "Daylight savings time flag," and "geopoint" columns from the dataframe and appended the result to the zip-code/neighborhood dataframe. This left us with a zip-code/neighborhood/latitude/longitude dataframe.

City of San Francisco COVID-19 Cases and Death Summarized by Geography Data:

Next, we needed to import and clean the COVID-19 data from the San Francisco COVID-19 by Geography and Date data source (link here:https://data.sfgov.org/COVID-19/COVID-19-Cases-by-Geography-and-Date/d2ef-idww) to then append to our zip-code/neighborhood/latitude/longitude dataframe (by zip code matching). When exporting the data, make sure you export it as a CSV file. We then imported this data into a pandas dataframe. We filtered the 'id' column for just zip code values that all contained the string '94'. Then, we grouped the dataframe by their common zip code value and calculated the Average New COVID-19 Cases per Day for a zip-code column. We then appended this dataframe to the zip-code/neighborhood/latitude/longitude dataframe via zip-code matching. The resultant dataframe, which we shall refer to as san_fran_metrics, contains the zip-code, neighborhood name, neighborhood latitude, neighborhood longitude, and average new COVID-19 cases per day, as shown in figure 1.

	Zip	Neighborhood	Latitude	Longitude	Average New COVID-19 Cases per Day
0	94102	Hayes Valley/Tenderloin/North of...	37.779329	-122.41915	3.769953
1	94103	South of Market	37.772329	-122.41087	2.890909
2	94107	Potrero Hill	37.766529	-122.39577	1.825472
3	94108	Chinatown	37.792678	-122.40793	0.447236
4	94109	Polk/Russian Hill (Nob Hill)	37.792778	-122.42188	3.140271
5	94110	Inner Mission/Bernal Heights	37.748730	-122.41545	7.716216
6	94112	Ingelside-Excelsior/Crocker-Amazon	37.720931	-122.44241	8.301802
7	94114	Castro/Noe Valley	37.758434	-122.43512	0.972477
8	94115	Western Addition/Japantown	37.786129	-122.43736	2.144796
9	94116	Parkside/Forest Hill	37.743381	-122.48578	0.854460
10	94117	Haight-Ashbury	37.770937	-122.44276	1.269767
11	94118	Inner Richmond	37.782029	-122.46158	0.990654
12	94121	Outer Richmond	37.778729	-122.49265	0.936652
13	94122	Sunset	37.758380	-122.48478	1.130233
14	94123	Marina	37.801028	-122.43836	0.911215
15	94124	Bayview-Hunters Point	37.732797	-122.39348	6.677273
16	94127	St. Francis Wood/Miraloma/West P...	37.734964	-122.45970	0.550725
17	94131	Twin Peaks-Glen Park	37.741797	-122.43780	0.781818
18	94132	Lake Merced	37.724231	-122.47958	1.101852
19	94133	North Beach/Chinatown	37.801878	-122.41018	0.731132
20	94134	Visitacion Valley/Sunnydale	37.719581	-122.41085	4.607477

Figure 2 Resultant Dataframe of San Francisco Values

Foursquare Wrangling:

We began wrangling the Foursquare wrangling data by passing the San Francisco neighborhood latitude and longitude values from the san_fran_metrics dataframe to the Foursquare's API. We requested the neighborhood name, latitude, longitude, venue name, latitude, longitude, and category from Foursquare's data warehouse. We

then grouped the venues by neighborhood (see bar chart in figure 2) and plotted them onto a San Francisco map using the Folium python package (Fig.3)

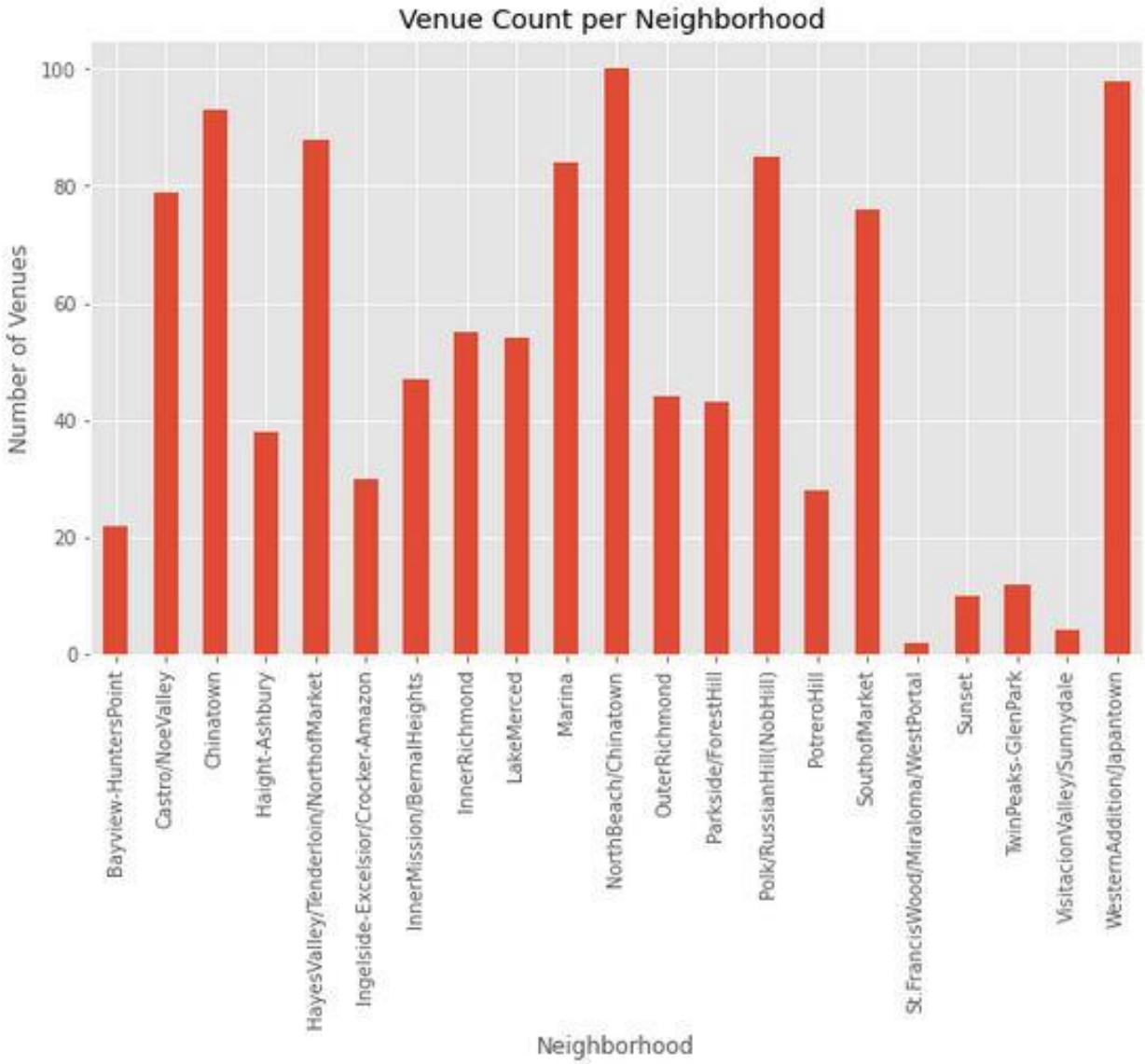

Figure 3 Venue Count Per San Francisco Neighborhood

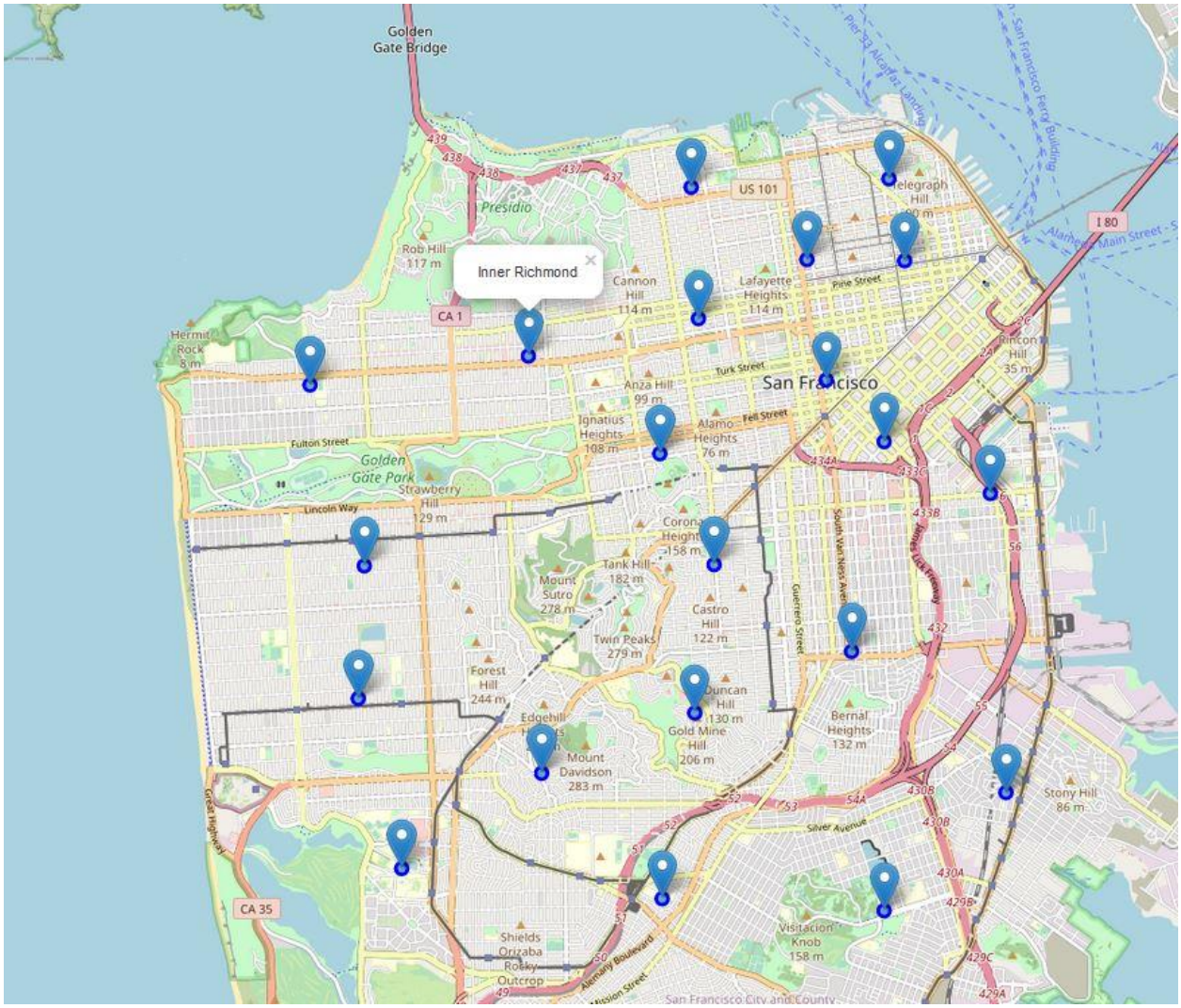

Figure 4 San Francisco Neighborhoods Plotted on San Francisco

Methodology

We now discuss the methodology where we describe the exploratory data analysis we used and how we utilized the k-means clustering machine learning algorithm to cluster neighborhoods by venue type and the relative risk of contracting COVID-19. We then categorized each cluster by their top three venue type by frequency and their relative risk of contracting COVID-19 as either "Low," "Medium," or "High".

We needed to perform one-hot encoding for all the venues and then group the one-hot encoded values by neighborhood and take the mean of each venue category's frequency of occurrence. We then appended the normalized average new COVID-19 cases per day value to the dataframe to get the dataframe shown in figure 5. We also plot the normalized average new COVID-19 cases per day in a bar chart in figure 6.

	Neighborhood	Average New COVID-19 Cases per Day	ATM	Accessories Store	Adult Boutique	African Restaurant	Alternative Healer
0	Hayes Valley/Tenderloin /North of...	0.423030	0.000000	0.000000	0.000000	0.000000	0.000000
1	South of Market	0.311115	0.000000	0.000000	0.000000	0.000000	0.000000
2	Potrero Hill	0.175469	0.000000	0.000000	0.000000	0.000000	0.000000
3	Chinatown	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
4	Polk/Russian Hill (Nob Hill)	0.342862	0.000000	0.000000	0.000000	0.023529	0.000000
5	Inner Mission/Bernal Heights	0.925446	0.000000	0.000000	0.000000	0.000000	0.000000
6	Ingelside-Excelsior/Crocker-Amazon	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000
7	Castro/Noe Valley	0.066871	0.000000	0.000000	0.012658	0.000000	0.000000
8	Western Addition/Japantown	0.216124	0.000000	0.000000	0.000000	0.000000	0.000000
9	Parkside/Forest Hill	0.051846	0.000000	0.000000	0.000000	0.000000	0.000000
10	Haight-Ashbury	0.104720	0.000000	0.026316	0.000000	0.000000	0.000000

Figure 5 Final Dataframe Used for K-means Clustering (only 10 of 20 Neighborhoods shown)

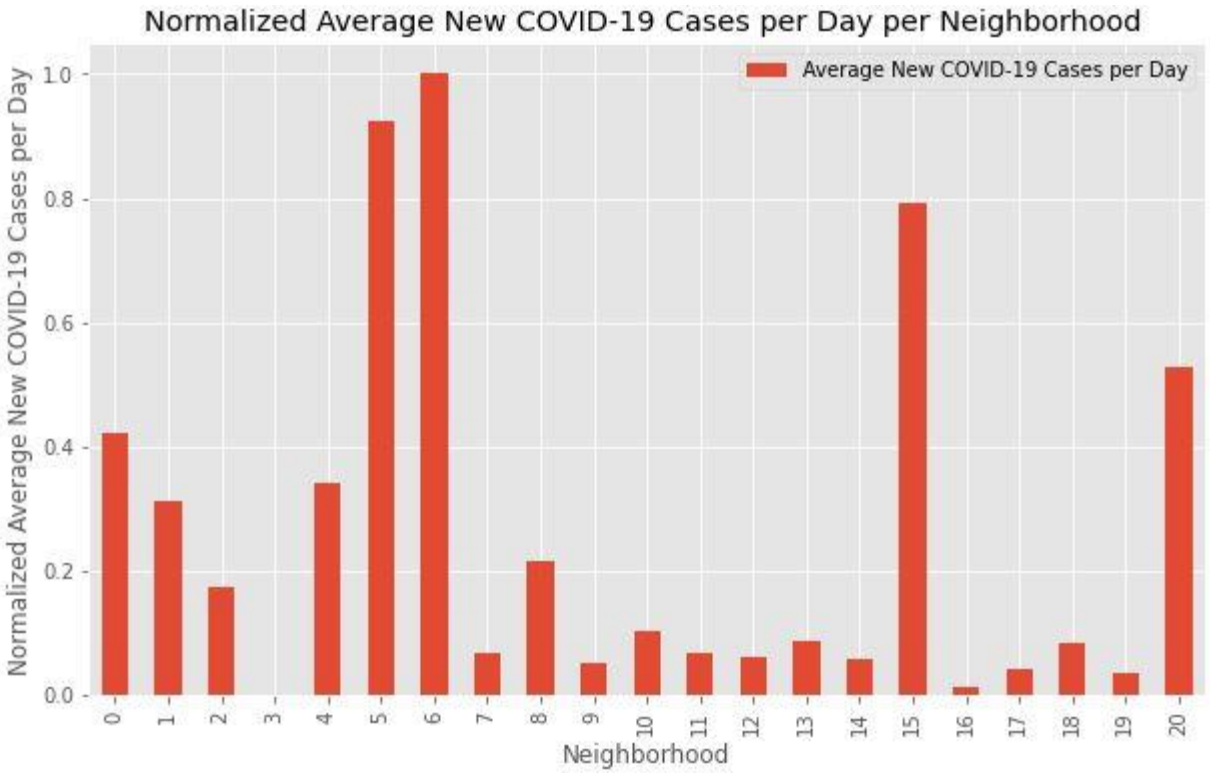

Figure 6 Normalized Average New COVID-19 Cases per Day per Neighborhood

Our next methodology task was to perform k-means clustering to cluster the neighborhoods into five clusters using the dataframe referred to in figure 5. We used k-means clustering to segment the neighborhoods based on what types of venues are commonly found in a neighborhood. We then created a new dataframe to append the cluster labels to the dataframe displayed in figure 7.

	Zip	Neighborhood	Latitude	Longitude	Average New COVID-19 Cases per Day	Cluster Labels	1st Most Common Venue	2nd Most Common Venue
0	94102	Hayes Valley/Tenderloin /North of...	37.779329	-122.41915	0.423030	0	Coffee Shop	Café
1	94103	South of Market	37.772329	-122.41087	0.311115	1	Nightclub	Cocktail Bar
2	94107	Potrero Hill	37.766529	-122.39577	0.175469	4	Breakfast Spot	Wine Shop
3	94108	Chinatown	37.792678	-122.40793	0.000000	0	Hotel	Bakery
4	94109	Polk/Russian Hill (Nob Hill)	37.792778	-122.42188	0.342862	4	Grocery Store	Gym / Fitness Center

Figure 7 K-Means Clustering Dataframe

Next, we utilized the dataframe in figure 7 to visualize the resulting five clusters utilizing folium, as shown in figure 8 below.

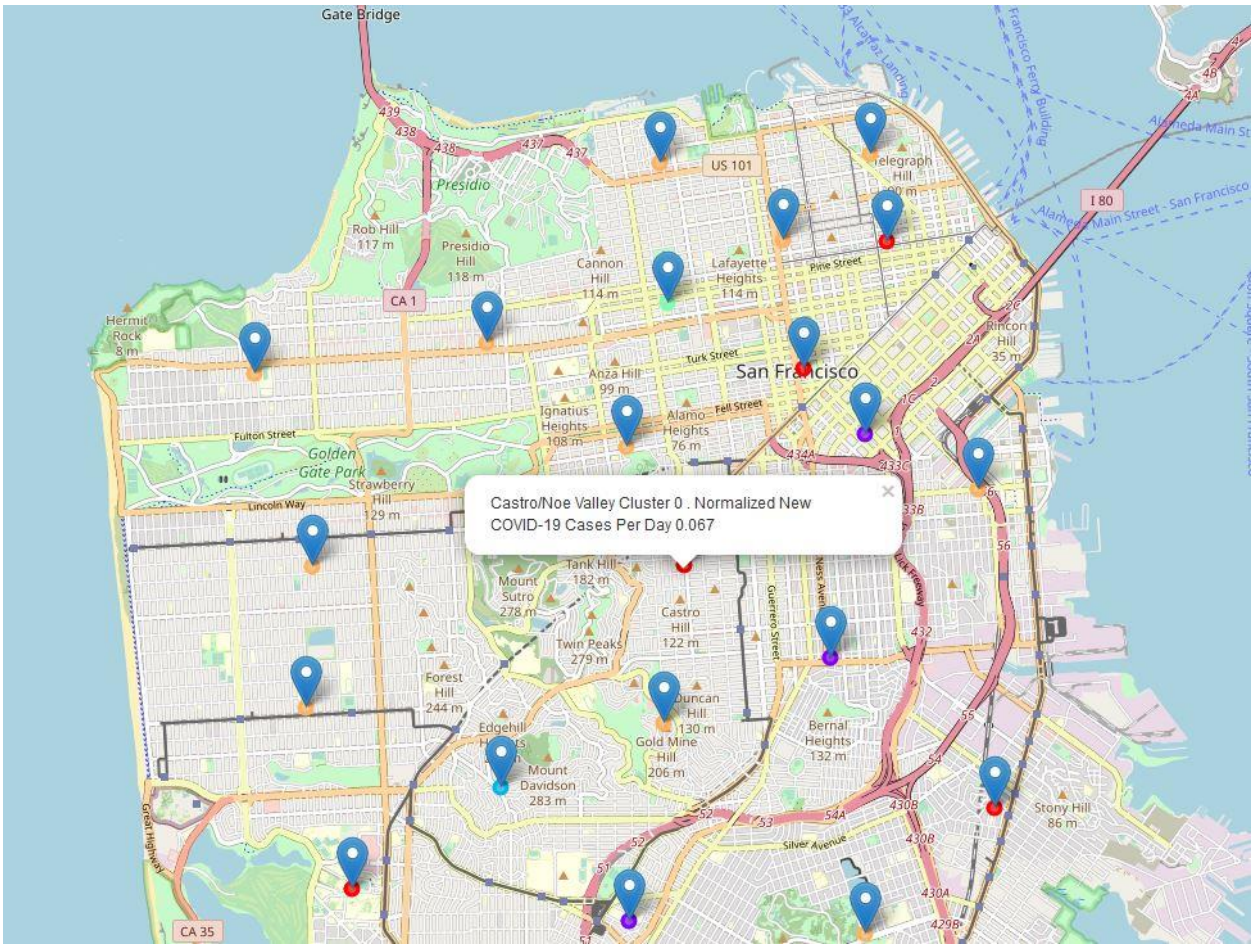

Figure 8 San Francisco Clustered Neighborhoods Plotted on San Francisco

We examined each cluster and determined the discriminating venue categories that distinguished each cluster. Based on the defining categories, we assigned a name to each cluster. We also defined a functioned that we used to categorize the relative risk of contracting COVID-19 in a venue's neighborhood as either "Low," "Medium," or "High" (see figure 9).

```
def covid_19_rank(val):
    if val <0.333:
        rank = 'Low'
    elif val<0.667:
        rank = 'Medium'
    else:
        rank = 'High'
    return rank
```

Figure 9 Relative Risk of Contracting COVID-19 Function

Results

In the results section, we summarize the results by analyzing each of the five clusters and classifying them based on their three most common venue types. We also attach the relative risk of contracting COVID-19 score to each neighborhood cluster by taking the mean of all the risk scores for neighborhoods within a cluster. We summarize the results by cluster (color) in the table below.

Table 1: Summary of Results

Cluster Number/Color	*Cluster Classification*	*COVID-19 Risk Score*
Cluster 0/Red	Coffee Shop, Café, and Hotel	Low
Cluster 1/Purple	Pizza Place, Mexican Restaurant, and Nightclub/Bar	High
Cluster 2/Light Blue	Bus Line, Trail, and Yoga Studio	Low
Cluster 3/Light Green	Bakery, Spa, and Café	Low
Cluster 4/Orange	Chinese Restaurant, Grocery Store, and Grocery Store	Low

Discussion

Taking a look at the results, it is clear that the relative risk of contracting COVID-19 in most venue neighborhoods is low. However, one could argue to avoid venues in cluster 1, the "Pizza Place, Mexican Restaurant, and Nightclub/Bar" due to its relatively high COVID-19 Risk Score. Also, if someone is looking to go to a coffee shop but are concerned about the risk of contracting COVID-19 in the area, the application provides them with a crude yet useful metric to gauge their risk. The utility of this analysis is clear, but it does have some limitations.

First of all, this analysis does not consider the potential risk reduction or increase that can occur by a user's own safety and health practices. Second of all, this application does not alert to whether a lockdown is in place or stores are closed. In future iterations of this application, we

would need to include this data into our analysis to better inform users on where they should shop to mitigate their risk of contracting COVID-19.

Conclusion

When utilizing location-based apps such as Google Maps, Lyft, or Uber, the application user does not know the relative risk of contracting COVID-19 in the restaurant's neighborhood they wish to eat at. This project addresses this problem by providing the user relative COVID-19 Risk Scores for different neighborhood clusters that contain different venues. For example, a user may want to get food from a pizza place and try to go to a restaurant in venue cluster 1. However, after seeing that the relative COVID-19 Risk Score is High, they may choose to order delivery from that restaurant instead of going in person. An application such as this helps inform consumer decisions about going to a restaurant or shopping in the age of COVID-19.