Identifying COVID-19 High Risk Neighborhoods in Boston

Raghurama Bukkarayasamudram

Background

Starting from February 2020, COVID-19 has swept the world in a very fast and ruthless fashion. What initially started off as a small news article from a relatively unknown part of China, COVID-19 has made its way all over the world. This posed a very unique situation to many of the people and businesses in the world- how to handle a pandemic. The closest example was the swine flu outbreak early in the decade, but that virus was nowhere near as deadly as the novel Coronavirus. Furthermore, there were existing vaccinations that people could take to be safe.

Currently, we are facing a far more dangerous virus- one that has no cure. This has impacted all levels of society from all over the world causing many disruptions in all sectors of life. From the obvious severe medical impact of battling a virus with no cure from many ill-equipped hospitals to economic shut down because businesses were forced to close, COVID-19 has had a significant impact that will forever change how the world will operate.

Problem

With the ongoing travel ban, closing of non essential businesses, limited reopening of essential businesses, and other restrictions, COVID-19 has severely hindered the economy, and it is more crucial than ever to play a role in halting the virus and spread. Obviously, the United States could have handled the virus better from a governmental standpoint which has underplayed the role of it several times, but there also exists a certain level of personal responsibility that people have to take to ensure that the spread of the virus is minimal or in ideal scenarios- non existent.

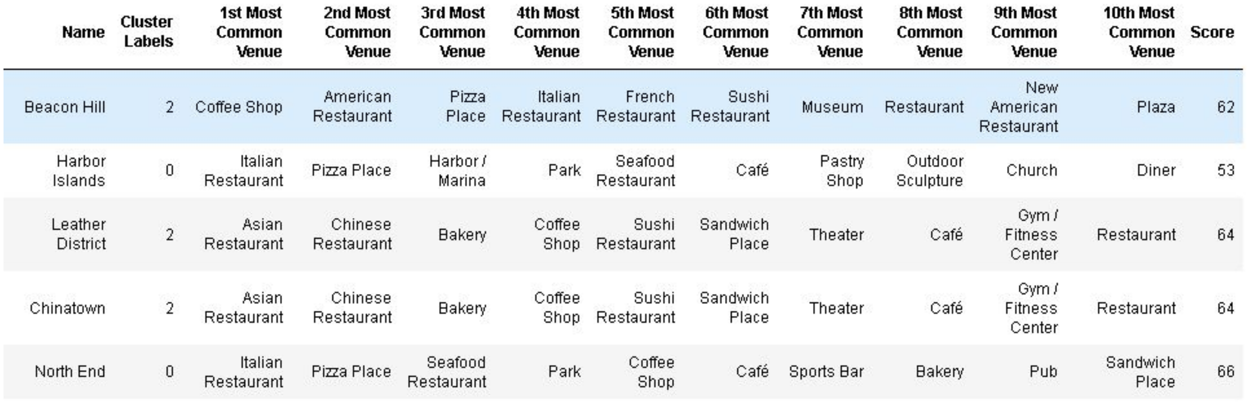
As there is no current cure to the virus, businesses have been set back rather harshly. However, in an increasingly connected world, it becomes easier for businesses to redeploy from an online section. While businesses can still operate, it becomes increasingly difficult for people to do the same. Performing everyday tasks still pose a dangerous threat- especially to the elderly or immuno-compromised demographics. As a result, it is very important to avoid areas of high risk or where exposure to the virus may be more common. My project's goal is to help people (and businesses) avoid high risk areas while still allowing themselves to be able to carry out daily and essential tasks. Essentially, I will be answering which neighborhoods to avoid during the pandemic.

Data and Methodology

To address this problem, I chose Boston as my city of choice to analyze. Boston is home to 20+ unique neighborhoods, and it has a variety of venues that make it one of the more diverse locations in the United States. With a great deal of venues in the city, also comes a great deal of risks- especially during the pandemic.

The data I used includes location data from various Boston neighborhoods. I initially created a dataframe that gave our zip codes and the neighborhood names, and combined it with another csv file that included the latitudes and longitudes of the zip codes. This combined data frame was then used to handle Foursquare API calls that collect neighborhood venue information.

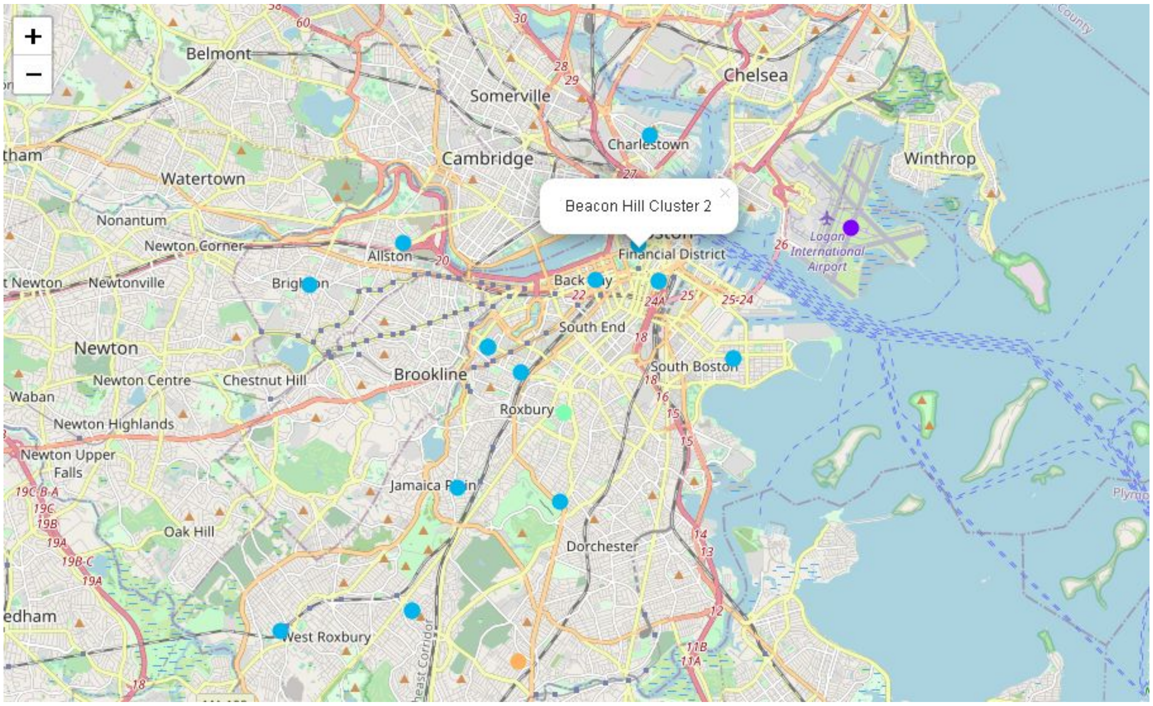
Foursquare API venue calls returned many unique categories of venues, but for the sake of simplicity, I needed to use an umbrella term to categorize each venue (i.e. “Italian Restaurant” would be categorized under a broader “Restaurant” term). I used the mean function to return the most common venues for each neighborhood, and calculated a score based on the types of venues. I calculated the score based on the risk factor associated with each venue from a scale of 1 to 10. Venues that involved close proximity or contact with one another were given a higher score to indicate a higher risk of COVID-19 transmission. Likewise, venues that were outdoors or involved less face to face interaction were given a lower score to indicate a smaller risk of exposure.



Name	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Score
Beacon Hill	2	Coffee Shop	American Restaurant	Pizza Place	Italian Restaurant	French Restaurant	Sushi Restaurant	Museum	Restaurant	New American Restaurant	Plaza	62
Harbor Islands	0	Italian Restaurant	Pizza Place	Harbor / Marina	Park	Seafood Restaurant	Café	Pastry Shop	Outdoor Sculpture	Church	Diner	53
Leather District	2	Asian Restaurant	Chinese Restaurant	Bakery	Coffee Shop	Sushi Restaurant	Sandwich Place	Theater	Café	Gym / Fitness Center	Restaurant	64
Chinatown	2	Asian Restaurant	Chinese Restaurant	Bakery	Coffee Shop	Sushi Restaurant	Sandwich Place	Theater	Café	Gym / Fitness Center	Restaurant	64
North End	0	Italian Restaurant	Pizza Place	Seafood Restaurant	Park	Coffee Shop	Café	Sports Bar	Bakery	Pub	Sandwich Place	66

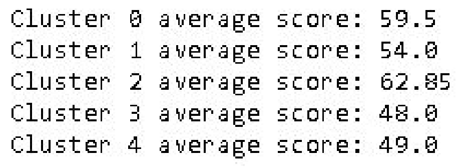
Once I had this information, I performed a K-means clustering algorithm on the data set using 5 centers. I found that after changing the number of iterations and centers, 5 was the optimal number of clusters that gave reliable information.

I used folium to display my data to make it more visually presentable. In this screenshot, it clearly segments all of the neighborhoods in Boston (some of them were omitted because they shared the same latitudes and longitudes). Each filled dot that shares a color is under the same cluster category.



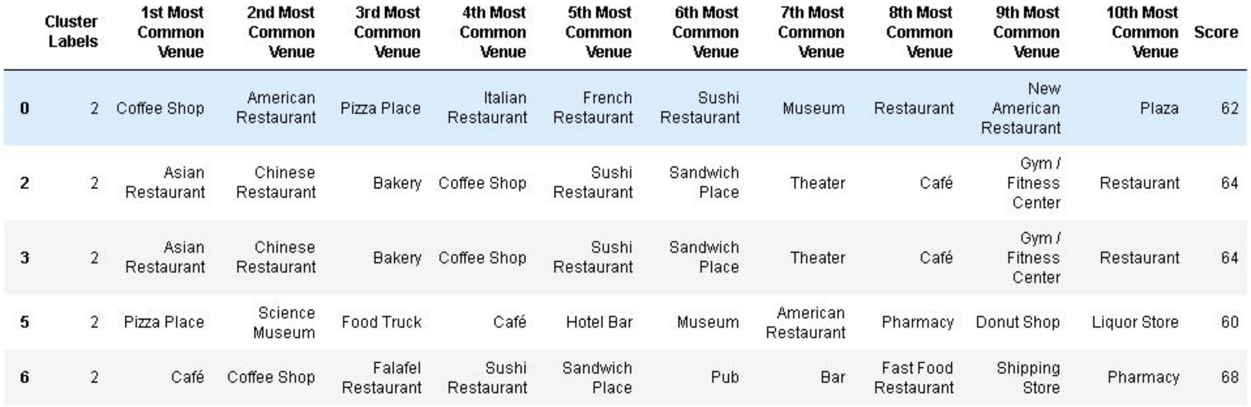
Results and Discussion

After k-means clustering was performed, it is evident that cluster 2 is by far the largest cluster, and coincidentally, it also seems to share the highest mean score out of all of the clusters. This is due to most of the venues in the neighborhoods in cluster 2 involving face-to-face contact or other forms of services that include close-proximity.



```
Cluster 0 average score: 59.5
Cluster 1 average score: 54.0
Cluster 2 average score: 62.85
Cluster 3 average score: 48.0
Cluster 4 average score: 49.0
```

When observing the types of common venues in cluster 2, it is evident that a restaurants and cafes are easily the most frequent. Since I assigned restaurants with a score of “7”, consistently, the mean COVID-19 score of neighborhoods in cluster 2 will be around 60-70.



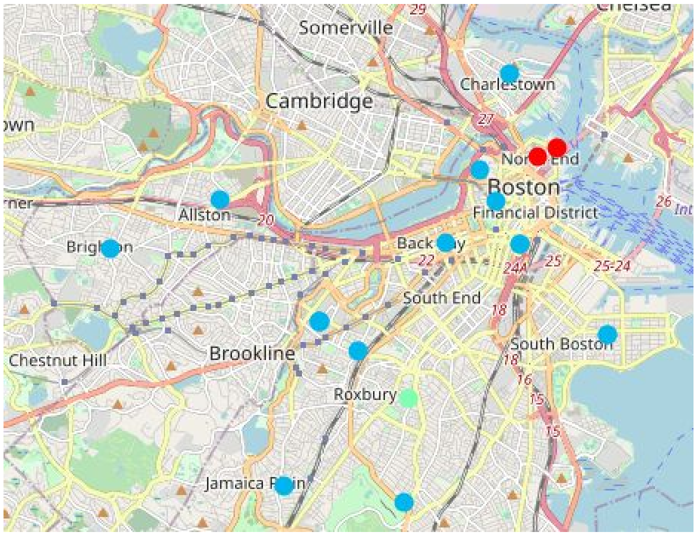
	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Score
0	2	Coffee Shop	American Restaurant	Pizza Place	Italian Restaurant	French Restaurant	Sushi Restaurant	Museum	Restaurant	New American Restaurant	Plaza	62
2	2	Asian Restaurant	Chinese Restaurant	Bakery	Coffee Shop	Sushi Restaurant	Sandwich Place	Theater	Café	Gym / Fitness Center	Restaurant	64
3	2	Asian Restaurant	Chinese Restaurant	Bakery	Coffee Shop	Sushi Restaurant	Sandwich Place	Theater	Café	Gym / Fitness Center	Restaurant	64
5	2	Pizza Place	Science Museum	Food Truck	Café	Hotel Bar	Museum	American Restaurant	Pharmacy	Donut Shop	Liquor Store	60
6	2	Café	Coffee Shop	Falafel Restaurant	Sushi Restaurant	Sandwich Place	Pub	Bar	Fast Food Restaurant	Shipping Store	Pharmacy	68

However, there are some limitations in the results that I obtained. While it identified the locations that are safer because of the venues, Foursquare API did not allow me to clearly identify exactly the type of data that I required to perform this analysis thoroughly. Perhaps, a different machine learning algorithm might have been more appropriate to answer the question that I posed since k-means clustering did not evenly cluster and segment the neighborhoods in Boston. If one cluster only contains two neighborhoods, the scoring values will not properly be taken into account.

In our case, the scoring values for our neighborhoods did match up with the cluster since the lowest value scores were in clusters 3 and 4.

Conclusion

After analyzing the results from the segmentation and clustering, I am able to answer the original problem that I posed. The neighborhoods in cluster 2 in Boston are more susceptible to COVID-19 exposure than the rest. Coincidentally, these neighborhoods also all tend to gather near the wealthier areas of Boston. These wealthy areas tend to have more venues such as restaurants, bars, etc that tend to spike the risk score. Essentially, to avoid the risk of exposure, people should perform essential tasks in neighborhoods like or near Roxbury.


Somerville
Cambridge
Charlestown
North End
Boston
Financial District
Allston
Brighton
Back Bay
South End
South Boston
Brookline
Chestnut Hill
Roxbury
Jamaica Plain