Comparing the Business Make Up of Zip Codes in Chicago Affected by COVID-19

Anthony Urgena

April 21, 2021

Table of Contents

Introduction

The COVID-19 pandemic has affected everyone globally. While there is still much to learn about the virus, we know that certain areas have been affected more than others. This applies to the city of Chicago, more specifically certain areas of the city itself have been disproportionally affected more than others. In this project we will gather data to determine the hardest and lowest hit areas by cases and by deaths. Business data of these locations will then be clustered to find similarities and differences between these areas of the city.

This report is targeted toward public health officials and city planners of Chicago as this could potentially help allocate resources accordingly. This could also help other public health officials and city planners of municipalities or other local and state government officials to aid with analysis of their own areas.

Data

Data Sources

COVID-19 data was gathered here from the city of Chicago data website. Per the website, "This site provides applications using data that has been modified for use from its original source, www.cityofchicago.org, the official website of the City of Chicago. The City of Chicago makes no claims as to the content, accuracy, timeliness, or completeness of any of the data provided at this site. The data provided at this site is subject to change at any time. It is understood that the data provided at this site is being used at one's own risk." The data is grouped by ZIP code, 59 with reported numbers. The data was grouped to the most recent week of data. This was then cleaned to include the ZIP Code with its population and location, cumulative cases, cases per 100 hundred thousand people, tests, tests per 100 thousand people, deaths, and deaths per 100 thousand people. Also gathered from the city of Chicago data website was a GeoJson file to draw the boundaries for the different zip codes.

Foursquare data was utilized to create a data frame of the 100 nearest businesses within 2.5 km from the location of the ZIP Code. Although some data overlaps, no duplicates were removed as they will be interpreted to still be inclusive of the area and is key to accurately cluster the location.

Data Cleaning

The first process was to make sure that all zip codes were populated. There was one unknown and the row was dropped as it would not be useful. Each zip code has a week start and a week end. To simplify, all rows were converted to date time format to help gather data from the most recent week. This was then put into a new working data frame. Any non-numbers would be replaced with zeroes. From this point, the only columns kept were

the zip code, cumulative numbers for cases, case rate, tests, test rate, deaths, death rate, population, and location. The location data provided was in a point object. To make it easier for processing later, these were converted to string objects and split into separate latitude and longitude columns.

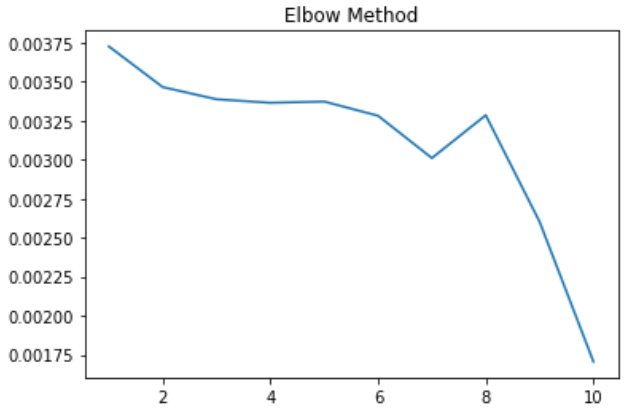
Data gathered from Foursquare came in a Json format and required flattening. The data was filtered down to venue name, category, latitude and longitude. A function was created to gather the nearby venues and appended to a list. One hot encoding was used to quickly create dummy variables for the different venues. This was then grouped by zip code by average occurrence. This produced 328 different venue categories. To get a quick overview of the five highest occurring business, a function was created to print out these results. A new data frame was created showing the top ten most common venues in the zip code to be added to the final data frame. A data frame without zip codes showing the average frequency of each venue category was created to prepare for cluster analysis.

Feature Selection

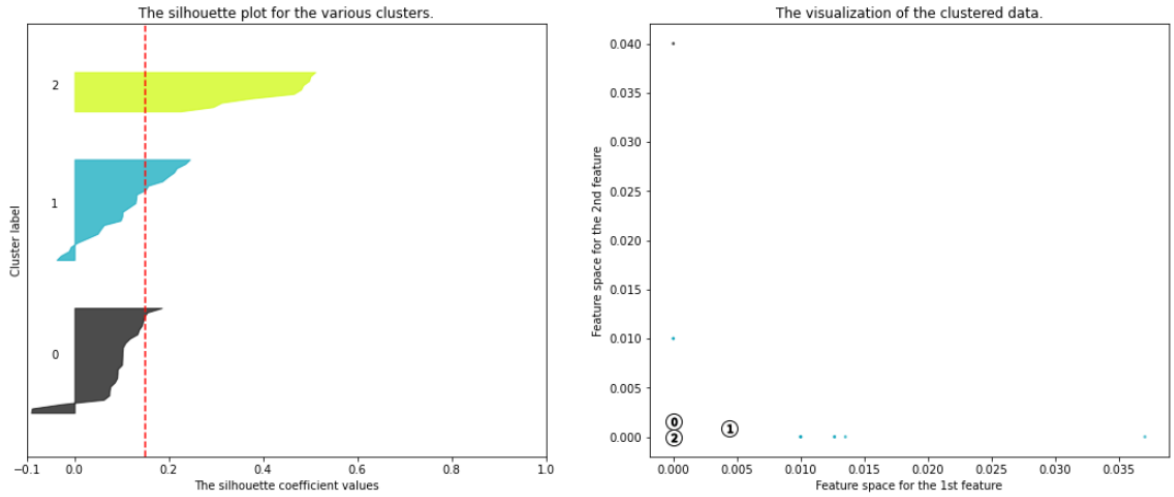
The data frame mentioned above will be clustered to help determine how similar the make up of the zip codes are to one another. This will be compared to the cumulative cases per one hundred thousand, deaths per one hundred thousand, and tests per one hundred thousand as provided by the city of Chicago. These rates will be utilized due to zip codes having different area size and population differences.

Methodology

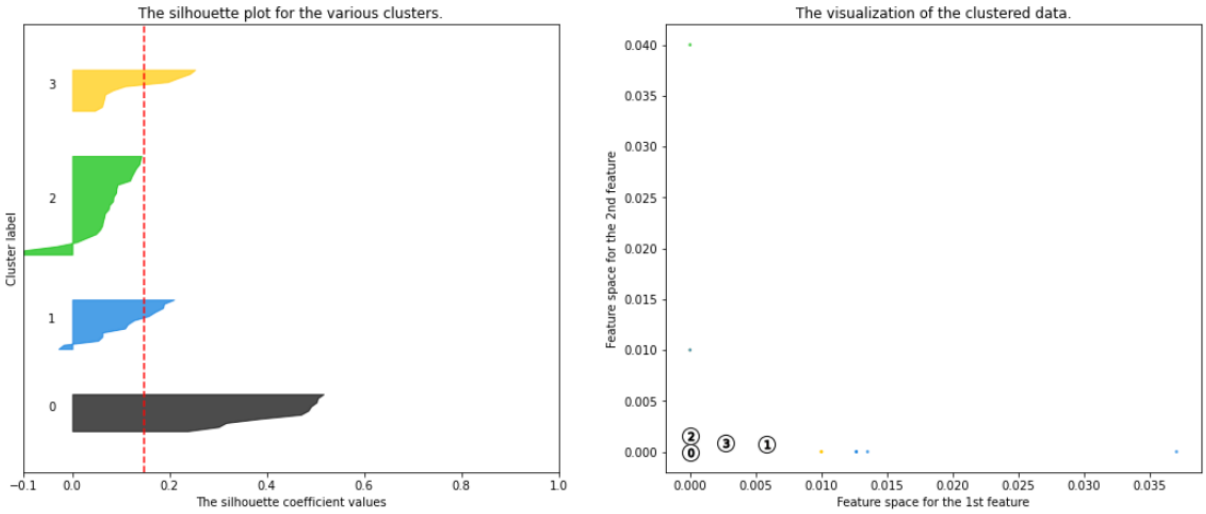
Since there was no predisposition on the make up of the business of the city the data above was clustered by k-means. This was used to help us discover unknown patterns and with these patterns we can then compare this with the case, death, and test rates to discover if there was a relationship. To help determine the best K, the elbow method and the silhouette method were both utilized. The elbow method shown below is not obvious but suggests k could be 3, 4, 5, or 6. The silhouette method was then used to help determine which would be the most suitable k. This tested between 2 and 10 clusters, although 2 would not be ideal and therefore excluded below, for suitability. Since the goal in using the silhouette method is to have equal sized rows and ideally no clusters below average, indicated by the dotted red line, only 3, 4, 5, and 6 will be displayed. In the end, 4 was chosen because it appeared to be the most uniform in terms of size and length.

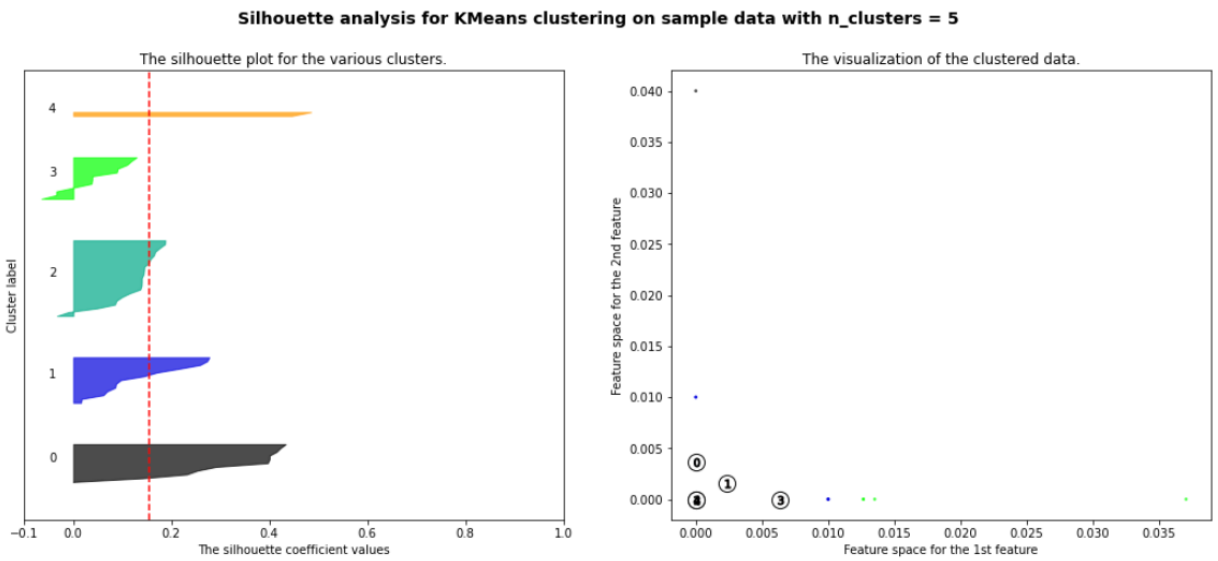


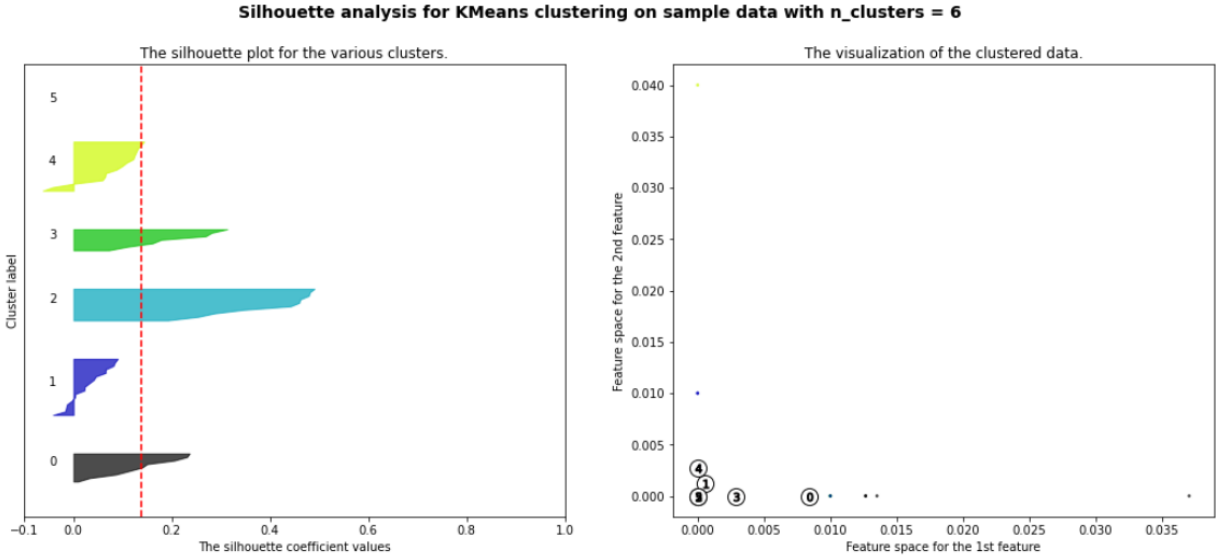
Silhouette analysis for KMeans clustering on sample data with n_clusters = 3



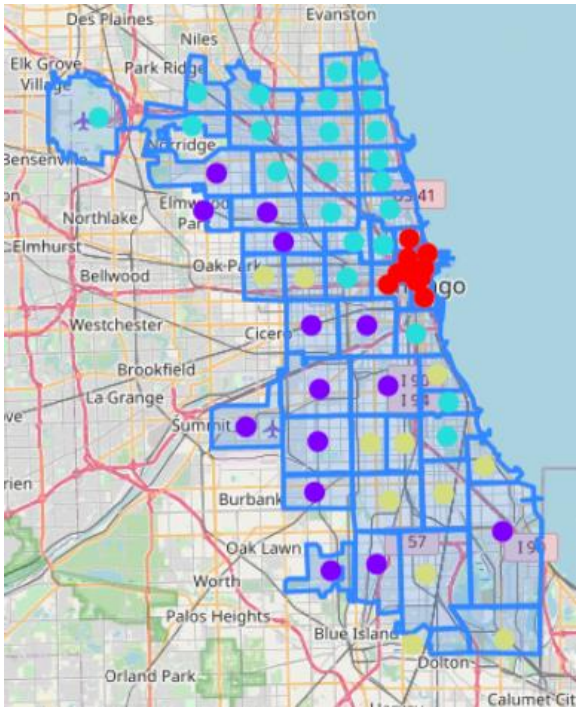
Silhouette analysis for KMeans clustering on sample data with n_clusters = 4



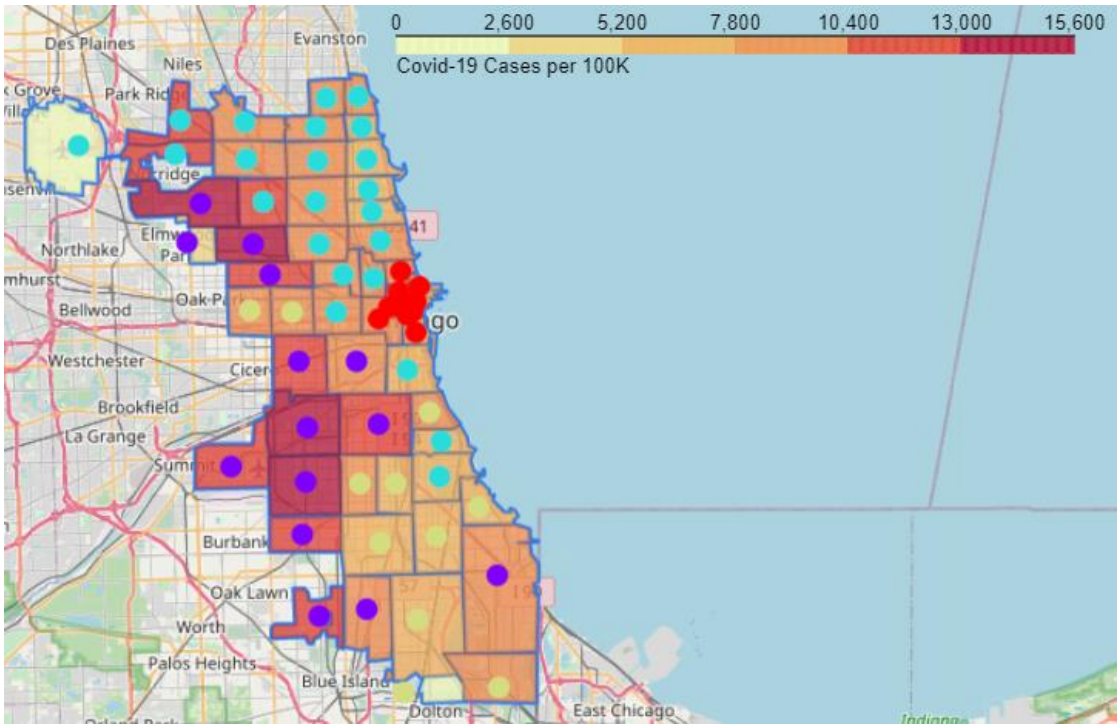


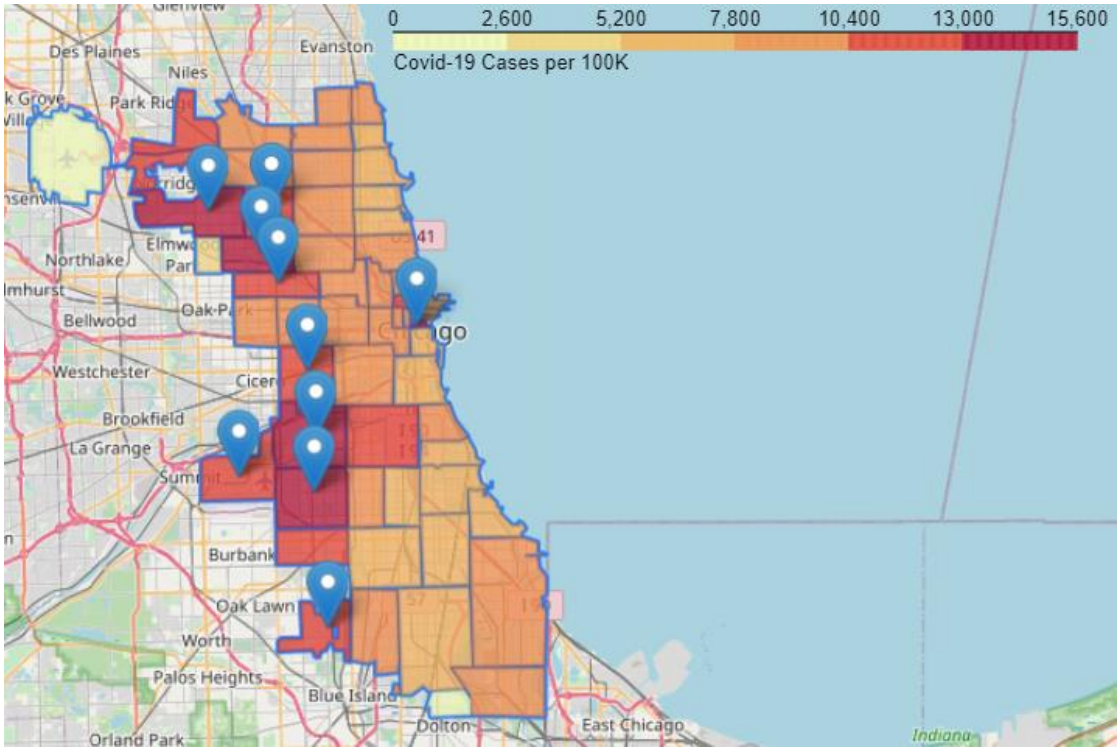


Now with k determined, the data was clustered and overlayed over the boundaries of the zip codes provided by the city of Chicago. It can be seen that there a few overlaps but for the most part it appears to be clustered geographically into a central, north, west, south, and far south. Cluster 1 is purple, cluster 2 is light blue, cluster 3 is yellow, and cluster 4 is the red. Using these clusters, choropleth maps were created for cases per one hundred thousand, deaths per one hundred thousand, and tests per one hundred thousand, one will be shown with the clusters and one with the location points of the top ten respective rates. These will be compared to see if there is a relationship between the cases, deaths, and tests against the clusters to determine if a pattern or other rationale can be extrapolated.



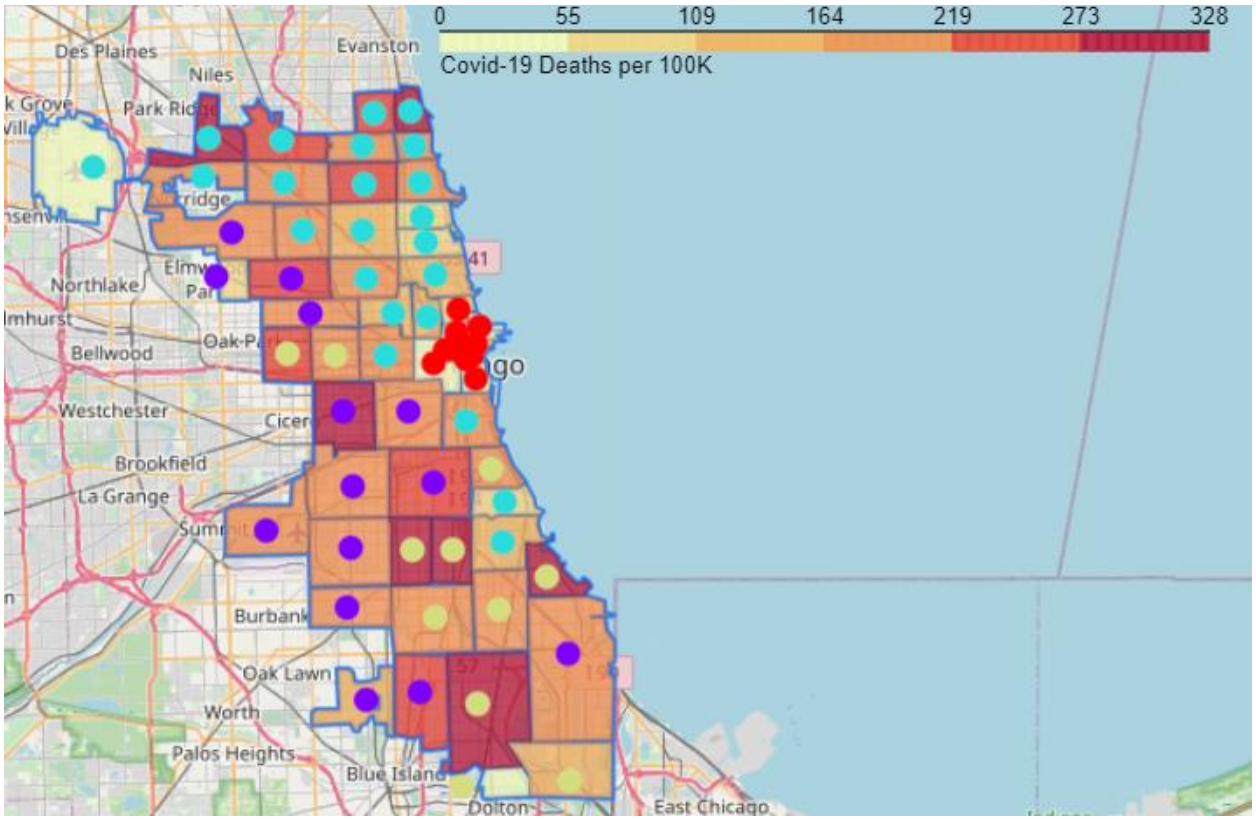
The first comparison is with the cases per one hundred thousand. When overlayed with the clusters, the maps shows that the hardest hit areas are mostly in cluster 1, purple. Looking at the ten highest case rates, there appears to be a concentration on the west side of the city with an outlier in the central area. Below that image is the data of the ten most common venues of the zip codes which will be discussed in a further section.

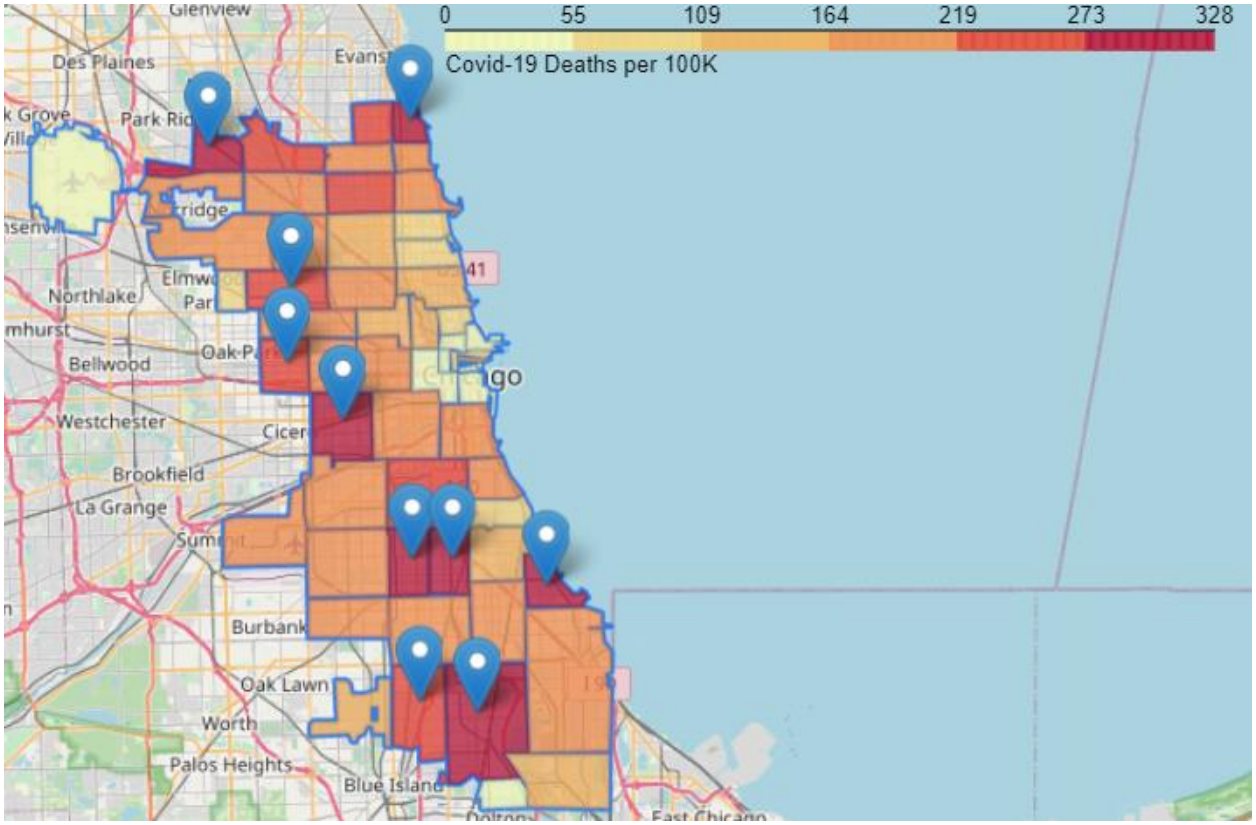




	ZIP Code	Cases per 100K	Population	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	60639	15600.4	90517	1	Mexican Restaurant	Donut Shop	Grocery Store	Fast Food Restaurant	Pharmacy	American Restaurant	Bank	Discount Store	Sandwich Place	Supermarket
1	60629	15018.3	111850	1	Mexican Restaurant	Discount Store	Sandwich Place	Pharmacy	Pizza Place	Fast Food Restaurant	Taco Place	Donut Shop	Video Store	Grocery Store
2	60632	14323.5	91039	1	Mexican Restaurant	Sandwich Place	Grocery Store	Bar	Taco Place	Discount Store	Pizza Place	Bank	Coffee Shop	Café
3	60634	13237.7	75995	1	Fast Food Restaurant	Mexican Restaurant	Clothing Store	Cosmetics Shop	Sandwich Place	Coffee Shop	Italian Restaurant	Bakery	Pizza Place	Deli / Bodega
4	60604	13171.4	782	4	Hotel	Park	Steakhouse	Theater	New American Restaurant	Italian Restaurant	Coffee Shop	Grocery Store	Bar	Mediterranean Restaurant
5	60638	12937.7	58797	1	Pizza Place	Rental Car Location	Mexican Restaurant	Convenience Store	Donut Shop	American Restaurant	Sandwich Place	Fast Food Restaurant	Bakery	Supermarket
6	60623	12385.6	85979	1	Mexican Restaurant	Sandwich Place	Discount Store	Taco Place	Bank	Pizza Place	Mobile Phone Shop	Donut Shop	Pharmacy	Gym / Fitness Center
7	60641	12314.3	71023	2	Mexican Restaurant	Coffee Shop	Park	Diner	Pizza Place	Asian Restaurant	Café	Sandwich Place	Bar	American Restaurant
8	60651	11990.3	63218	1	Mexican Restaurant	Donut Shop	Discount Store	Sandwich Place	Grocery Store	Fried Chicken Joint	Fast Food Restaurant	Gas Station	Pharmacy	Video Store
9	60655	11703.2	28804	1	Pizza Place	Mexican Restaurant	Sandwich Place	Bar	Italian Restaurant	Pharmacy	Fast Food Restaurant	Breakfast Spot	Pub	Convenience Store

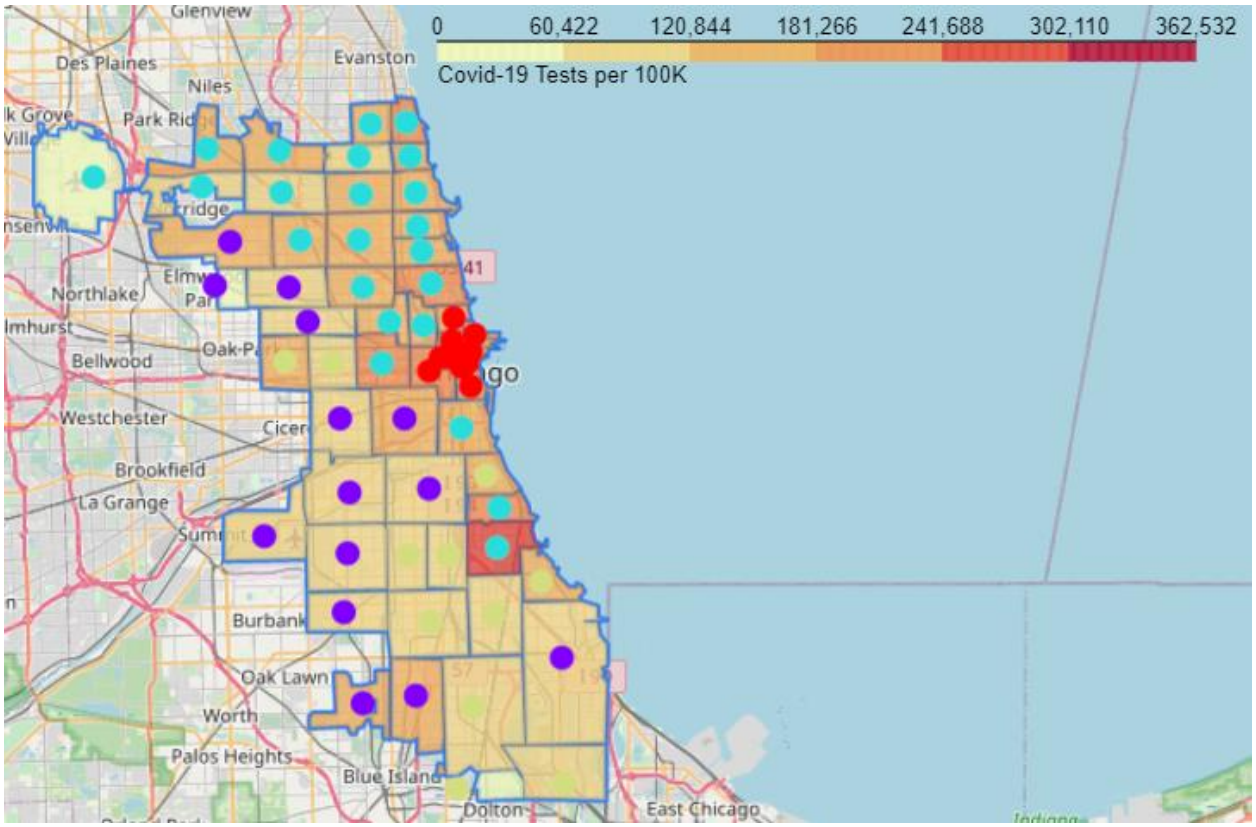
The second comparison is with the deaths per one hundred thousand. When overlayed with the clusters, the map does not show that any one cluster has been affected than another. Looking at the ten highest death rates, second below, there does not appear to be an obvious pattern except that they are not located in the central area. Below that image is the data of the ten most common venues of the zip codes of the highest death rates to be discussed later.

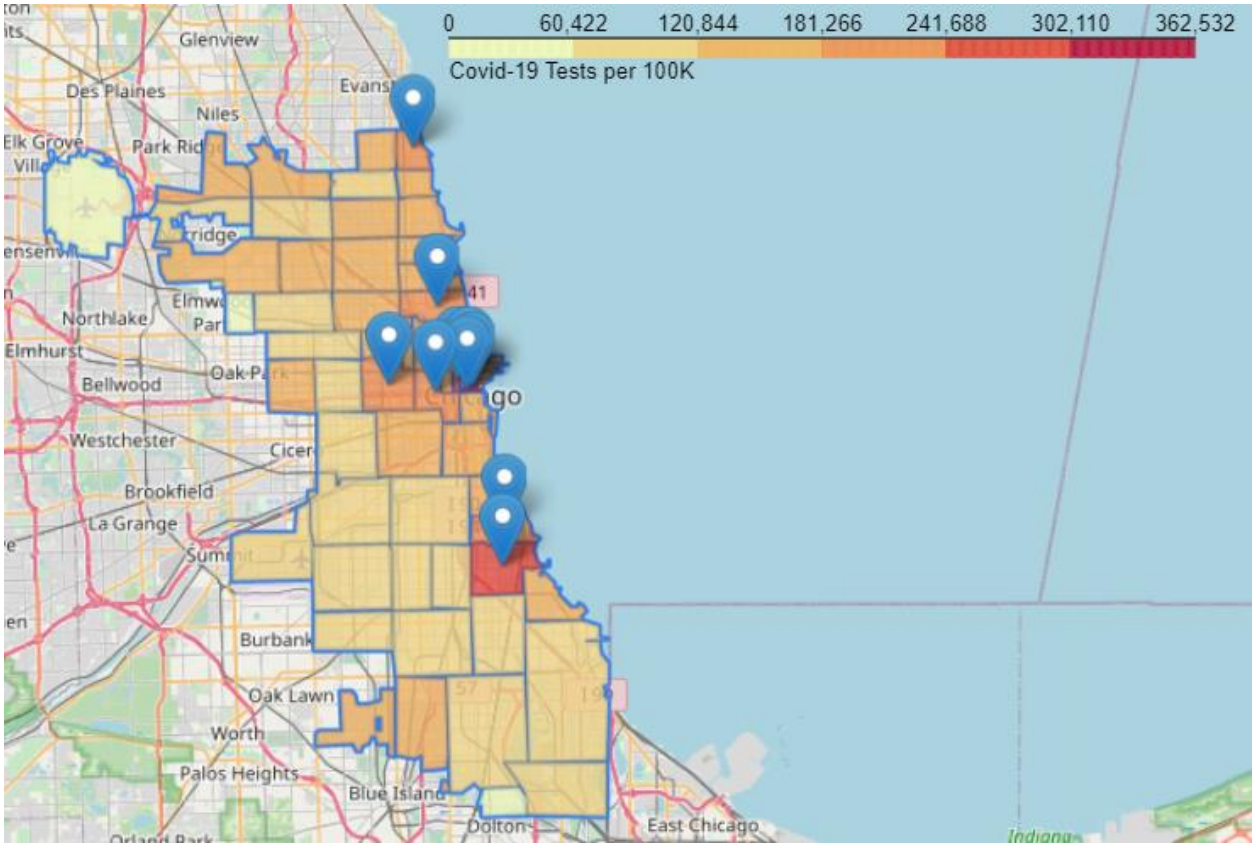

0
55
109
164
219
273
328
Covid-19 Deaths per 100K
Des Plaines
Evanston
Niles
Park Ridge
Northlake
Bellwood
Westchester
Brookfield
La Grange
Burbank
Oak Lawn
Worth
Palos Heights
Blue Island
Dolton
East Chicago


0
55
109
164
219
273
328
Covid-19 Deaths per 100K
Glenview
Des Plaines
Park Ridge
Northlake
Bellwood
Westchester
Brookfield
La Grange
Burbank
Oak Lawn
Worth
Palos Heights
Blue Island

	ZIP Code	Deaths per 100K	Population	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
0	60649	328.1	46024	3	Discount Store	Fast Food Restaurant	Bank	Pizza Place	Sandwich Place	Fried Chicken Joint	Park	Chinese Restaurant	Harbor / Marina	Supermarket
1	60626	319.7	49730	2	Park	Beach	Mexican Restaurant	Coffee Shop	Café	Pizza Place	Grocery Store	Gym	Thai Restaurant	Indian Restaurant
2	60631	314.9	29529	2	Italian Restaurant	American Restaurant	Coffee Shop	Salon / Barbershop	Pizza Place	Mexican Restaurant	Park	Sushi Restaurant	Grocery Store	Pharmacy
3	60636	304.3	32203	3	Fast Food Restaurant	Discount Store	Grocery Store	Mexican Restaurant	Donut Shop	Sandwich Place	American Restaurant	Fried Chicken Joint	Pharmacy	Park
4	60628	292.2	66724	3	Sandwich Place	Fast Food Restaurant	Fried Chicken Joint	Discount Store	Liquor Store	Pharmacy	Grocery Store	Train Station	Donut Shop	Chinese Restaurant
5	60623	288.4	85979	1	Mexican Restaurant	Sandwich Place	Discount Store	Taco Place	Bank	Pizza Place	Mobile Phone Shop	Donut Shop	Pharmacy	Gym / Fitness Center
6	60621	275.5	29042	3	Fast Food Restaurant	Gas Station	Sandwich Place	Fried Chicken Joint	American Restaurant	Discount Store	Grocery Store	Donut Shop	Seafood Restaurant	Train Station
7	60639	253.0	90517	1	Mexican Restaurant	Donut Shop	Grocery Store	Fast Food Restaurant	Pharmacy	American Restaurant	Bank	Discount Store	Sandwich Place	Supermarket
8	60643	250.7	49870	1	Pizza Place	Sandwich Place	Bar	Pharmacy	Donut Shop	Grocery Store	Fried Chicken Joint	Gas Station	Breakfast Spot	Brewery
9	60644	249.4	47712	3	Discount Store	Park	Seafood Restaurant	Donut Shop	Fast Food Restaurant	ATM	Breakfast Spot	Fried Chicken Joint	Sandwich Place	Record Shop

The final comparison is with the tests per one hundred thousand. When overlayed with the clusters, the map does not show that any one cluster has been testing more than another. Looking at the ten highest test rates, second below, the concentration appears to be in the central area. Below that image is the data of the ten most common venues of the zip codes of the highest test rates to be discussed later.





	ZIP Code	Tests per 100K	Population	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
0	60604	362532.0	782	4	Hotel	Park	Steakhouse	Theater	New American Restaurant	Italian Restaurant	Coffee Shop	Grocery Store	Bar	Mediterranean Restaurant
1	60637	271543.0	47454	2	Science Museum	Park	Coffee Shop	Bookstore	Café	Pizza Place	History Museum	Grocery Store	Caribbean Restaurant	Spa
2	60606	266075.5	3101	4	Hotel	Italian Restaurant	New American Restaurant	Bar	Coffee Shop	Mediterranean Restaurant	Mexican Restaurant	Steakhouse	Grocery Store	Theater
3	60626	234456.1	49730	2	Park	Beach	Mexican Restaurant	Coffee Shop	Café	Pizza Place	Grocery Store	Gym	Thai Restaurant	Indian Restaurant
4	60612	212241.0	34311	2	Restaurant	Brewery	Mexican Restaurant	Breakfast Spot	Coffee Shop	Dive Bar	Bar	Pizza Place	Park	Grocery Store
5	60607	203362.5	29591	4	Italian Restaurant	Hotel	Coffee Shop	New American Restaurant	Yoga Studio	Sandwich Place	Grocery Store	Pizza Place	Park	Ice Cream Shop
6	60603	201873.9	1174	4	Hotel	Italian Restaurant	Park	Bar	Theater	Coffee Shop	Mediterranean Restaurant	Steakhouse	Grocery Store	Donut Shop
7	60602	190594.9	1244	4	Hotel	Italian Restaurant	Park	Bar	Theater	Steakhouse	Coffee Shop	Seafood Restaurant	Mediterranean Restaurant	Grocery Store
8	60614	188095.3	71308	2	Pizza Place	Coffee Shop	Grocery Store	Park	Liquor Store	Gym	Bakery	Sushi Restaurant	Seafood Restaurant	Italian Restaurant
9	60615	183369.8	41563	2	Coffee Shop	Park	Science Museum	Art Gallery	Beach	BBQ Joint	Café	History Museum	Pizza Place	Breakfast Spot

Results

Looking at the maps and data frames above, it appears that there is a visual correlation in terms of highest cases to cluster 1. Looking at the data frame for the ten highest rates, eight are in cluster 1. In all of the cluster 1 ZIP codes a Mexican restaurant is either the first, second, or third most common.

It appears inconclusive in terms of deaths. Visually, the ten highest death rates are not concentrated in any particular area. Upon inspecting the data, it is fairly even between clusters 1, 2, and 3.

For the ten highest test rates, it appears to be centrally located with an outlier to the far north side of the city and to the south.

Discussion

When analyzing the clusters, I was not surprised in how the clusters were formed. From my personal experience, the Chicago neighborhoods have their own characteristics and when talking to other Chicagoans it can be broken down geographically to the loop, central area, north, south, and west sides. I was surprised at how clearly tied the highest case rates were to cluster 1. It can be explored more deeply on the make-up of the neighborhood by using socioeconomic and other demographic information. As the scope of this report is on the business make up of the ZIP code these considerations would have to be put to the side for another project.

This would be especially helpful in determining the pattern for the death rate as there was no clear cluster with increased death rates. To help research this another project using health care access data may be applicable. In viewing the test rates, it is obvious that the concentration is in the central or loop area. Looking at the make up of the business data, hotels, theaters, and other tourist attractions are a large part of the area. The population of residents is also significantly lower. This could be broken down by researching the professions of those living in this area.

My recommendations based on these results would be to increase testing to the Western part of the city. I would also recommend for city planners to promote more spaces for social distancing as well as for providing more vaccine availability for the residents. I would recommend prioritizing this area over others because if the spread of the virus could be slowed down then the death rate should also fall down and can then allocate resources based on more recent number.

Conclusion

In this study, business data was gathered from Foursquare to help relate the business make up of the different ZIP codes of Chicago. This data was then clustered for comparison against the city's data for cumulative case, death, and test rates. These were then compared to the areas of the city with the highest Covid-19 case, death, and test rates. It was discovered that one of the clusters had a strong relation to the ten highest case rates. Although the death and test rates were not as obvious the visualization could help in allocating resources from a city public health or urban planning perspective.

Future Directions

Although out of the scope of this project, these visualizations could be improved when combined with socioeconomic and other access to health care data. This project focused on the business make up of the different areas of Chicago which is only one small aspect of public health and urban planning. As discussed above, this could be a building block for helping allocate resources. With data about available spaces, city planners would have the opportunity about where distribution of municipal resources could be most effective.