Introduction

The effects of and each area's reaction to the COVID-19 pandemic have been far from homogenous. As cases in the United States surge, it is urgently important that we better understand why this virus affects some populations more than it does others. The volume of public data and timeliness of the subject makes this an apt project for data science. Using available data, we can determine how the popular locations of each area are related to that area's COVID-19 burden.

The problem of not fully understanding how this coronavirus spreads is a pressing matter to local, state, national, and global health departments. Uncovering how hyperlocal (per zip code) differences in popular venue categories are affecting COVID-19 cases can help inform health department guidelines and help predict future hotspots. Because this project only deals with zip codes within Alameda County, it would be of immediate importance to their health department but expanding the scope beyond one county would enhance the statistical power.

Data

	Zip Code	Cases per 100,000 People	Latitude	Longitude
0	94501	211.334113	47.771396	18.101184
1	94502	124.158516	37.735143	-122.241527
2	94505	0.000000	37.891224	-121.616664
3	94514	0.000000	37.861842	-121.626666
4	94536	257.043353	37.560459	-121.973594

Table 1: Five entries from the database compiled from the data from Alameda Public Health

I used three datasets to elucidate the relationship between confirmed COVID-19 cases and popular local venues in Alameda county. COVID-19 rates are hosted, and updated daily, by Alameda Public Health at: https://services3.arcgis.com/1iDJcsklY3l3KIjE/arcgis/rest/services/AC_Rates_Zip_Code/FeatureServer/0/query?where=1%3D1&outFields=*&outSR=4326&f=json. For each of the 53 zip codes in Alameda county, this dataset contains the city, COVID-19 case count, population, and COVID-19 case rate. The provided zip codes can be converted to latitude and longitude by GeoPy for the FourSquare API request (Table 1). My analysis only requires the zip code and case rate from this dataset.

The next dataset is the result of using the FourSquare Places API to find recommended venues in each zip code. FourSquare API requests return recommended venues in each zip code and their GPS coordinates. From these data, I created a new pandas data frame (Table 2).

To display the data as a choropleth map, I required a GeoJSON file containing the outlines of each zip code within Alameda County. I chose the GeoJSON San Francisco "Bay Area Zip Codes" which is hosted by DataSF (https://data.sfgov.org/Geographic-Locations-and-Boundaries/Bay-Area-ZIP-Codes/u5j3-svi6).

Methodology

Data visualization allows for efficient digestion of large datasets. To this end, I first used the folium Python library to create a choropleth map of Alameda County to visualize the differences in case rates between zip codes as differences in colors or shading (Figure 1). This map requires the GeoJSON file with outlines of each Bay Area zip code and the case rates per zip code under “features” in the Alameda Public Health dataset. Because both datasets include zip codes, I used zip codes as the key to pair case rates with zip code outlines. When plotting the choropleth and all later maps, I chose Castro Valley, California as an approximate geographical center of Alameda County. I then displayed a choropleth of "COVID-19 Cases in Alameda County per 100,000 people" with case rates increasing as zip code shades morph from light yellow into dark green (Figure 1).

My next step was to pair cases of the new coronavirus with popular venues. Beginning with case rates and zip codes from the Alameda Public Health dataset, I found corresponding GPS coordinates using the GeoPy Python library. With these coordinates, I had enough data for calls to FourSquare’s Places API. This requires a “GET” request to the FourSquare API with the “explore” endpoint. For this study, I used a venue radius of 600m for each zip code.

	Zip Code	Zip Code Latitude	Zip Code Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	94501	47.771396	18.101184	KRTKOVANIE BÁNSKY	47.769601	18.105973	Home Service
1	94501	47.771396	18.101184	Nástupište 4	47.772500	18.094763	Platform
2	94502	37.735143	-122.241527	La Val's Pizza	37.737610	-122.241001	Italian Restaurant
3	94502	37.735143	-122.241527	La Penca Azul	37.737667	-122.240767	Mexican Restaurant
4	94502	37.735143	-122.241527	Coffee and Tea Traders	37.737493	-122.240265	Coffee Shop

Table 2: First 5 rows of a table containing the most popular venues in each zip code and their coordinates.

To work with venue categories, I converted categories to numbers using one hot encoding. This created a new data frame where the value for the correct venue category is 1 and the venue for all other categories is 0. From here, I grouped venue categories by zip code with fractional values representing the frequency at which each category is recommended. Then by ordering each venue by frequency, I created a data frame containing the ten most common venues per zip code (Table 3).

	Zip Code	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	94501	Home Service	Platform	Ethiopian Restaurant	Flower Shop	Flea Market	Fish Market	Filipino Restaurant	Fast Food Restaurant	Farmers Market	Farm
1	94502	ATM	Japanese Restaurant	Playground	Deli / Bodega	Park	Coffee Shop	Sandwich Place	Shipping Store	Shopping Mall	Mexican Restaurant
2	94505	Mexican Restaurant	Cosmetics Shop	Diner	Pharmacy	Sushi Restaurant	Fast Food Restaurant	Tanning Salon	Salon / Barbershop	Bank	Sandwich Place
3	94536	Pool	Lake	Gym / Fitness Center	Middle Eastern Restaurant	Flower Shop	Flea Market	Fish Market	Filipino Restaurant	Fast Food Restaurant	Farmers Market
4	94538	Grocery Store	Bagel Shop	Bakery	Ice Cream Shop	Juice Bar	Breakfast Spot	Sushi Restaurant	Furniture / Home Store	Bubble Tea Shop	Fried Chicken Joint

Table 3: First 5 rows of a table containing the top 10 most common popular venues in each zip code.

With the top venue categories per zip code, I could begin grouping Alameda zip codes based on popular venues using k-means clustering. I chose 3 clusters because additional clusters contained only 1 zip code and 12 iterations of the *k*-means algorithm to insure optimal inertia. To visualize the clustering, I plotted markers in each zip code on a map with a different color marker for each cluster (Figure 2). I then compared COVID-19 case rates per cluster using a box plot created using pyplot from the matplotlib Python library (Figure 3).

The final map (Figure 4) is a combination of the first two. A choropleth map with markers in each zip code corresponding to its cluster and zip code shades morph from light yellow into dark red as COVID-19 case rates rise.

Results

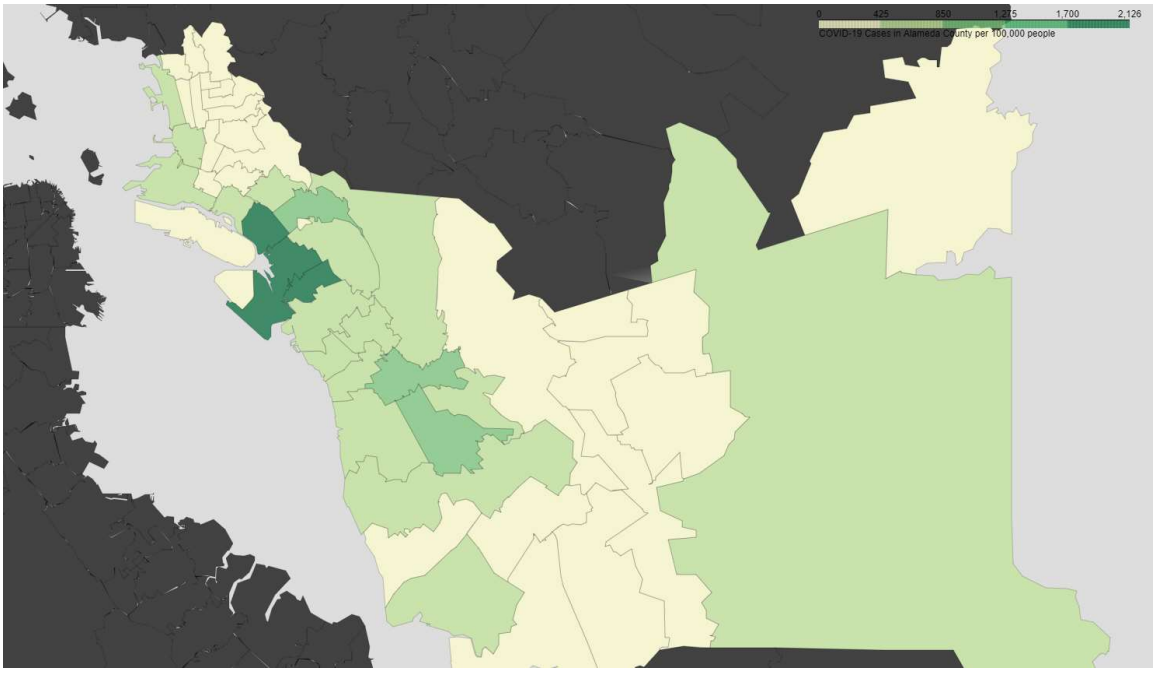

Figure 1: Choropleth map of Alameda County. COVID-19 case rates represented by color.

The choropleth map (Figure 1) indicates that COVID-19 cases are highest (darkest green) in a few zip codes representing the most urban areas of Alameda County. More inland areas seem to be dealing with fewer cases of the new coronavirus, however.

k-means clustering of the zip codes by popular venues resulted in three clusters. Cluster 0 (Table 5) contains 41 zip codes with shops and restaurants. Cluster 1 (Table 6) contains 2 zip codes that feature playgrounds and zoos and cluster 2 (Table 7) contains 4 zip codes with popular trails and parks.

	Zip Code	Cases per 100,000 People	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
10	94545	660.303465	Hobby Shop	Trail	Zoo Exhibit	Electronics Store	Flower Shop	Flea Market	Fish Market	Filipino Restaurant	Fast Food Restaurant	Farmers Market
23	94586	0.000000	Trail	Zoo Exhibit	Electronics Store	Flower Shop	Flea Market	Fish Market	Filipino Restaurant	Fast Food Restaurant	Farmers Market	Farm
38	94618	128.691398	Trail	Gym / Fitness Center	Tunnel	Baseball Field	Zoo Exhibit	Electronics Store	Flea Market	Fish Market	Filipino Restaurant	Fast Food Restaurant
47	94708	119.012236	Trail	Park	Hobby Shop	Garden	Electronics Store	Flea Market	Fish Market	Filipino Restaurant	Fast Food Restaurant	Farmers Market

Table 5: 10 most common venues of zip codes in Cluster 0

	Zip Code	Cases per 100,000 People	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
8	94542	609.639027	Playground	Wine Bar	Zoo Exhibit	Electronics Store	Flea Market	Fish Market	Filipino Restaurant	Fast Food Restaurant	Farmers Market	Farm
40	94621	2020.219730	Playground	Zoo Exhibit	Ethiopian Restaurant	Flower Shop	Flea Market	Fish Market	Filipino Restaurant	Fast Food Restaurant	Farmers Market	Farm

Table 6: 10 most common venues of zip codes in Cluster 1

	Zip Code	Cases per 100,000 People	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	94501	220.522553	Home Service	Platform	Ethiopian Restaurant	Flower Shop	Flea Market	Fish Market	Filipino Restaurant	Fast Food Restaurant	Farmers Market	Farm
1	94502	124.158516	ATM	Japanese Restaurant	Playground	Deli / Bodega	Park	Coffee Shop	Sandwich Place	Shipping Store	Shopping Mall	Mexican Restaurant
2	94505	0.000000	Mexican Restaurant	Cosmetics Shop	Diner	Pharmacy	Sushi Restaurant	Fast Food Restaurant	Tanning Salon	Salon / Barbershop	Bank	Sandwich Place
4	94536	272.700817	Pool	Lake	Gym / Fitness Center	Middle Eastern Restaurant	Flower Shop	Flea Market	Fish Market	Filipino Restaurant	Fast Food Restaurant	Farmers Market
5	94538	356.310944	Grocery Store	Bagel Shop	Bakery	Ice Cream Shop	Juice Bar	Breakfast Spot	Sushi Restaurant	Furniture / Home Store	Bubble Tea Shop	Fried Chicken Joint
7	94541	955.619737	Mexican Restaurant	Neighborhood	Food & Drink Shop	Convenience Store	Deli / Bodega	Department Store	Flower Shop	Flea Market	Fish Market	Filipino Restaurant
11	94546	503.159179	Park	Ice Cream Shop	Korean Restaurant	Ramen Restaurant	Food Truck	Market	Theater	Bubble Tea Shop	Hawaiian Restaurant	Grocery Store
12	94550	425.736249	Pool	Mobile Phone Shop	Theater	Park	Bed & Breakfast	Ethiopian Restaurant	Fish Market	Filipino Restaurant	Fast Food Restaurant	Farmers Market
13	94551	518.426668	Construction & Landscaping	Zoo Exhibit	Ethiopian Restaurant	Flower Shop	Flea Market	Fish Market	Filipino Restaurant	Fast Food Restaurant	Farmers Market	Farm
15	94555	140.170325	Farm	Business Service	Coffee Shop	Park	Zoo Exhibit	Fabric Shop	Flower Shop	Flea Market	Fish Market	Filipino Restaurant

Table 7: First 10 rows of the 10 most common venues of zip codes in Cluster 2

When plotting clusters onto a map, Cluster 2 (red dots, Figure 2) clearly dominates with its 41 zip codes. Cluster 0 (green dots, Figure 2) zip codes are close to the forested areas.

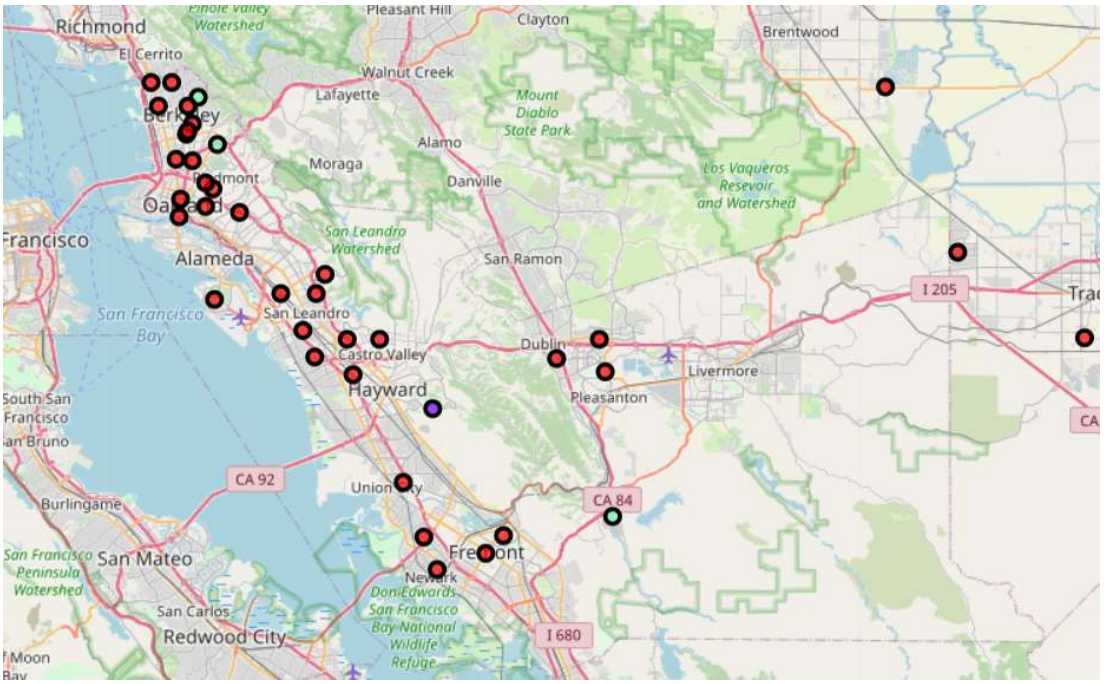

Figure 2: Map of Alameda County with the 3 zip code clusters represented by colored markers.

Only having a few zip codes in 2 of the 3 clusters makes statistical comparisons between clusters difficult. From the box plot (Figure 3), Clusters 0 and 1 are not different statistically. Cluster 2 case rates could be the lowest but since it contains only 2 zip codes, no conclusions can be made. Graphing both clusters and case rates (Figure 4) also shows no obvious relationship between clusters and COVID-19 cases.

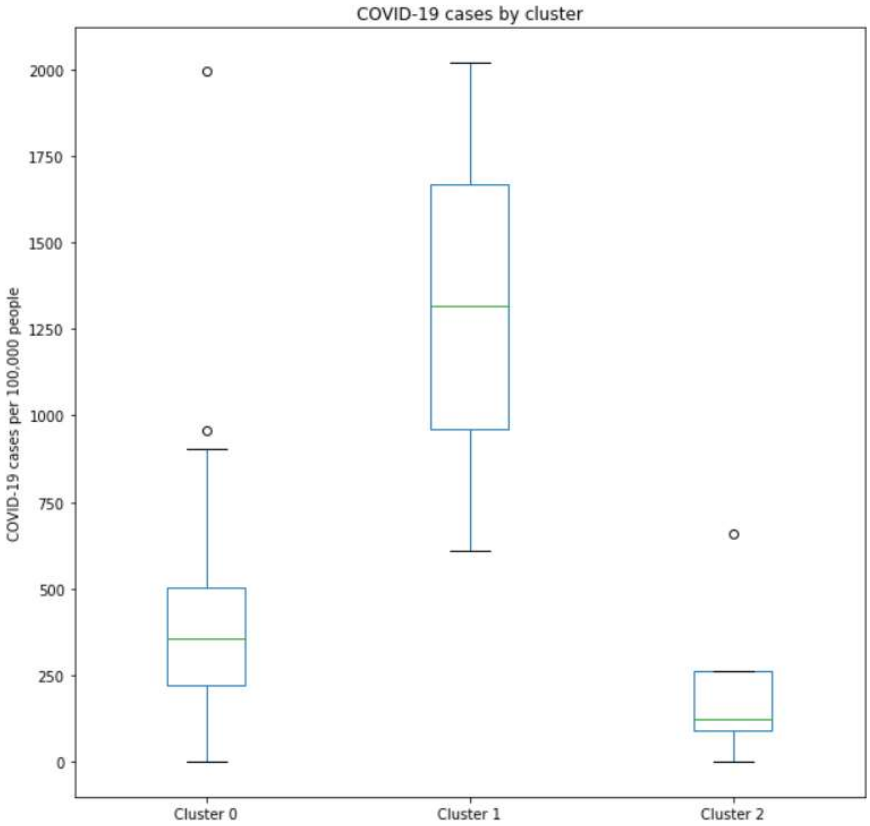

Figure 2: Box plot of COVID-19 case rates per cluster.

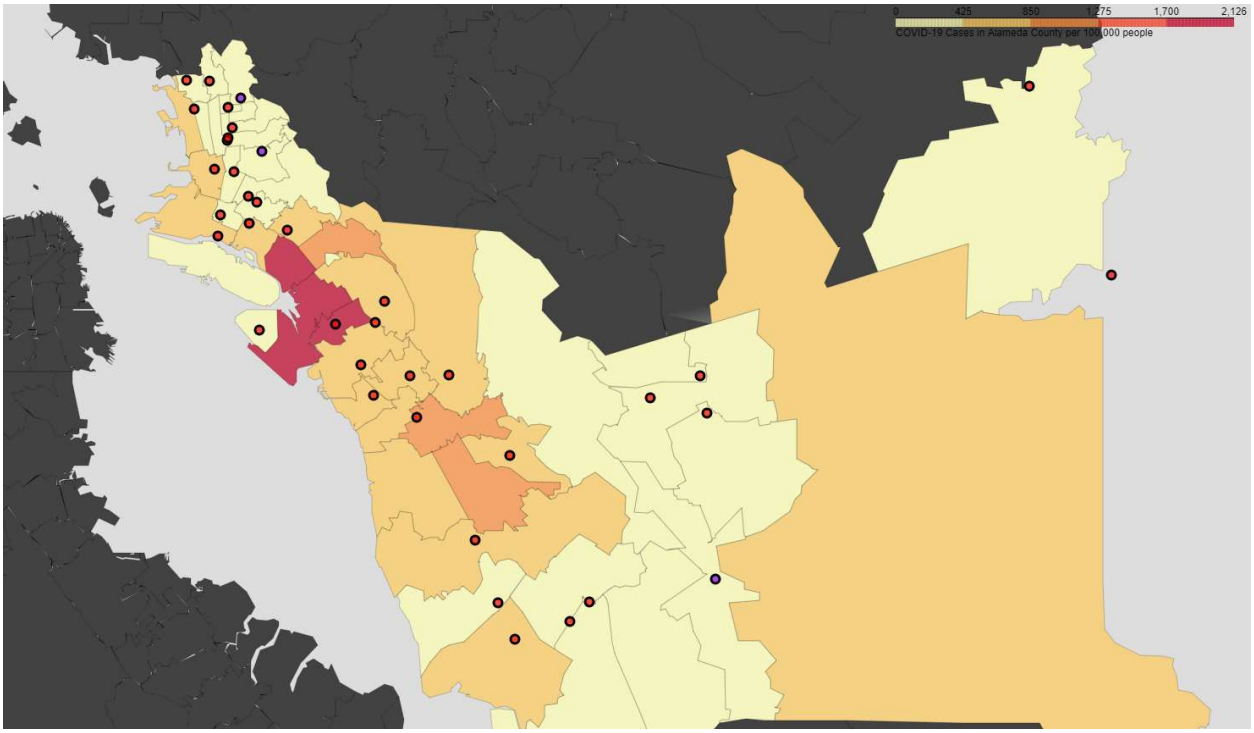

Figure 3: Choropleth map of Alameda County COVID-19 case rates with colored markers representing clusters.

Discussion

The results from this study are inconclusive. There could be a relationship between popular venue categories and COVID-19 cases but the data from Alameda County are not enough to demonstrate it. Since almost the entire county clusters into one group, I do not believe that its venues vary enough for any differences to be detected. The most obvious solution is to repeat the study but include data from the entire San Francisco Bay Area or California. Creating my own, more precise venue categories might reveal interesting trends that the previous categories were too broad to show. Also, it would be best if we could divide the county with a grid, randomly sample the map, then classify the samples by zip code. This would allow us to more accurately label zip codes even though Alameda zip code areas are irregularly shaped.

I would similarly be interested in looking into whether increased numbers of venue visits increase cases of the new coronavirus. Maybe the type of venue does not matter as much as its popularity.

As a capstone project, this study went great! Before this IBM professional certificate, I knew little about data science and analysis using Python. For this project, I had the opportunity to demonstrate my proficiency in using Jupyter Notebook, working with databases in SQL and Python, machine learning, making interactive and informative maps, and plotting results using Python.

Conclusion

Combining FourSquare location data with COVID-19 data is still promising for future epidemiological studies. This study was not powerful enough to find a link between where people like to go and their likelihood of testing positive for COVID-19 but it might only require a larger dataset. With more data comes less noise; we should encourage our local governments to continue hosting and regularly updating their online public health datasets.