Impact of Neighborhood Composition on COVID-19 Rates in Los Angeles

Matthew Morris

March 28, 2021

1. Introduction/Business Problem

COVID-19 had a profound impact on the way businesses operate in the US and elsewhere in the world. In order to see how COVID might continue to affect different neighborhoods where I may want to open a business, and or how another pandemic may affect a newly opening business, I'd like to determine if there is a correlation between neighborhood composition in terms of restaurants, outdoor activities, and economic specialties and the impact of COVID-19 on those neighborhoods. I will do a clustering very similar to the one we did for Toronto and New York, then add a dimension of COVID rates and overall population to determine if certain neighborhoods were disproportionately affected, and if so, which ones.

Beyond determining the impact for a potential new business, this report may be useful for municipalities or local governments preparing for pandemic scenarios in the future, to determine the highest likelihood areas of spread and which activities may need to be limited first. It may also indicate who the most at risk people and businesses are - and hence what kinds of businesses and areas need to close or distance first.

Finally, this report may begin to shed light on how successful lockdowns were at preventing COVID in urban areas. If there is a lower incidence of COVID in urban areas that are typically crowded, it may be worth digging into what percentage of all venues stayed open during COVID.

2. Data

1.1 Data Sources.

The three main data sources necessary for this report will be:

1) Data on COVID case rates by neighborhood. This report will use data publicly available on the LA county website, and only focus on LA counties. http://publichealth.lacounty.gov/media/coronavirus/locations.htm
2) Geographical data about each neighborhood. For the geographical coordinates of those neighborhoods, I will rely on USC's neighborhood data for social change, located here, which can be exported as a CSV: https://usc.data.socrata.com/dataset/Los-Angeles-Neighborhood-Map/r8qd-yxsr

3) Neighborhood composition data (ie, the kinds of venues in each neighborhood and how frequently they occur). Given the sets of geographical data above, I use the FourSquare API to pull venues per neighborhood and determine the composition.

1.2 Data Processing.

I will only be focused on neighborhoods in Los Angeles as opposed to the surrounding suburbs and exurbs. I will simply filter neighborhood data from the public health website for LA county on neighborhoods containing "Los Angeles."

As features for the models themselves, I will extract the neighborhood composition of each neighborhood in terms of the types of venues using data from the FourSquare API. In order to do this, a set of venue types will be pulled for the latitude and longitude of each LA neighborhood. Those venue types will then be one-hot encoded and averaged among all venue types in a neighborhood. This will give data for us to be able to cluster similar neighborhoods by common venue types (composition).

The COVID case rate will then be examined per neighborhood cluster to see if any clusters have a higher prevalence of covid than others.

3. Methodology

I will use multiple rounds of k-means clustering to determine the composition of certain neighborhoods by venue type. I ended up doing two rounds of clustering for this examination, since Exploratory data analysis revealed that the first round of clustering merely found outliers, neighborhoods on the outskirts of LA that were either wealthy or suburban.

For the first round of clustering, I tried many different values for k in our k-means machine learning algorithm. What I determined was that this first round of clustering essentially did outlier detection, distinguishing urban city areas from suburban or exurban areas. This left me with many of the urban neighborhoods we were trying to distinguish from one another in a single group, and separate groups that were only single points defining the substantially different (and different from each other) neighborhoods in the hills and outskirts of LA.

To better answer the question at hand, I removed these suburban and exurban clusters (outliers) and used the one large cluster for all urban areas to dig in further. Doing this, we got better resolution defining how urban areas varied and how those variations may have impacted COVID case rates.

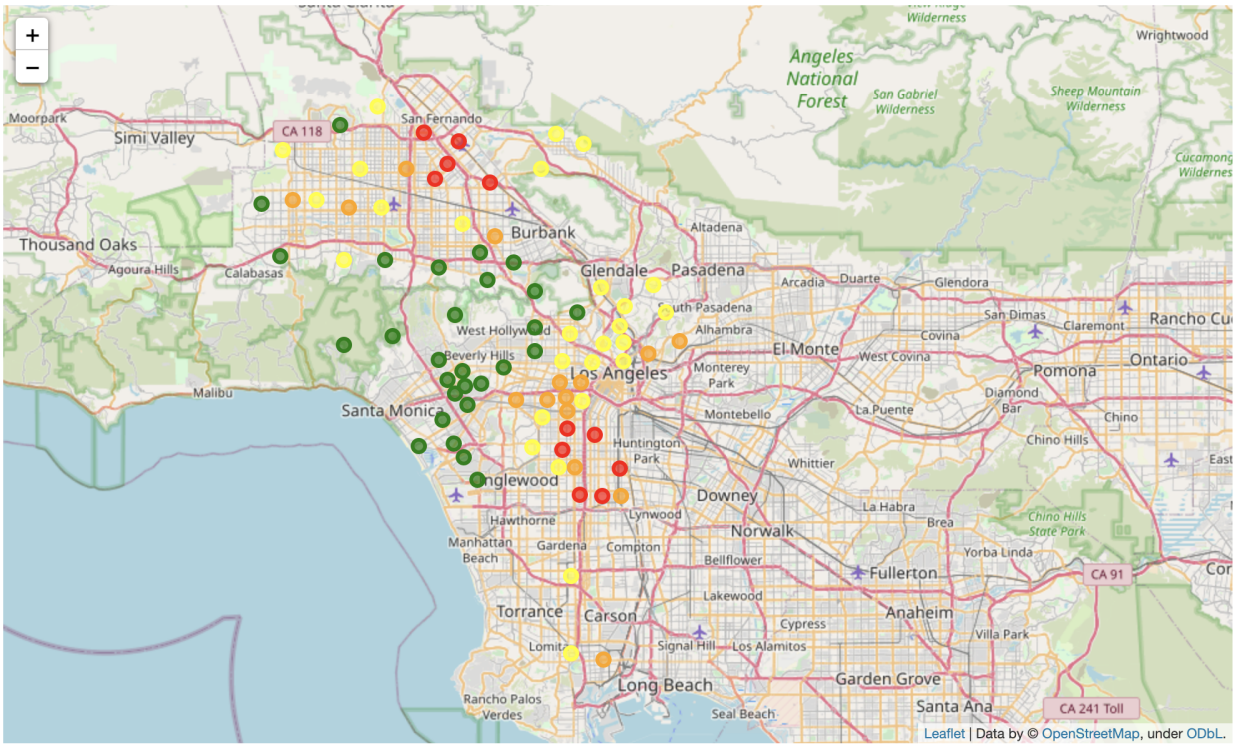
Once we had this more granular data, I was able to get a clearer picture of how neighborhood composition affected case rates. We examined the 2 clusters with the highest average cases/100k residents, and how they differed from the 3 clusters with lower COVID rates using tables and bar charts.

4. Exploratory Data Analysis

3.1 Quickly Visualize COVID Case Rates by Neighborhood

In order to get a first impression of where there was a high prevalence of COVID geographically in LA, we visualized this on a map using Folium. Green data points had relatively low incidence of COVID; red had high incidence (Green, Yellow, Orange, Red - in order of highest concentration of cases per 100k residents). As you can see on this map, there are several neighborhood clusters with higher prevalence of COVID than others; let's see if this relates to any particular neighborhood composition population density, or other factors:

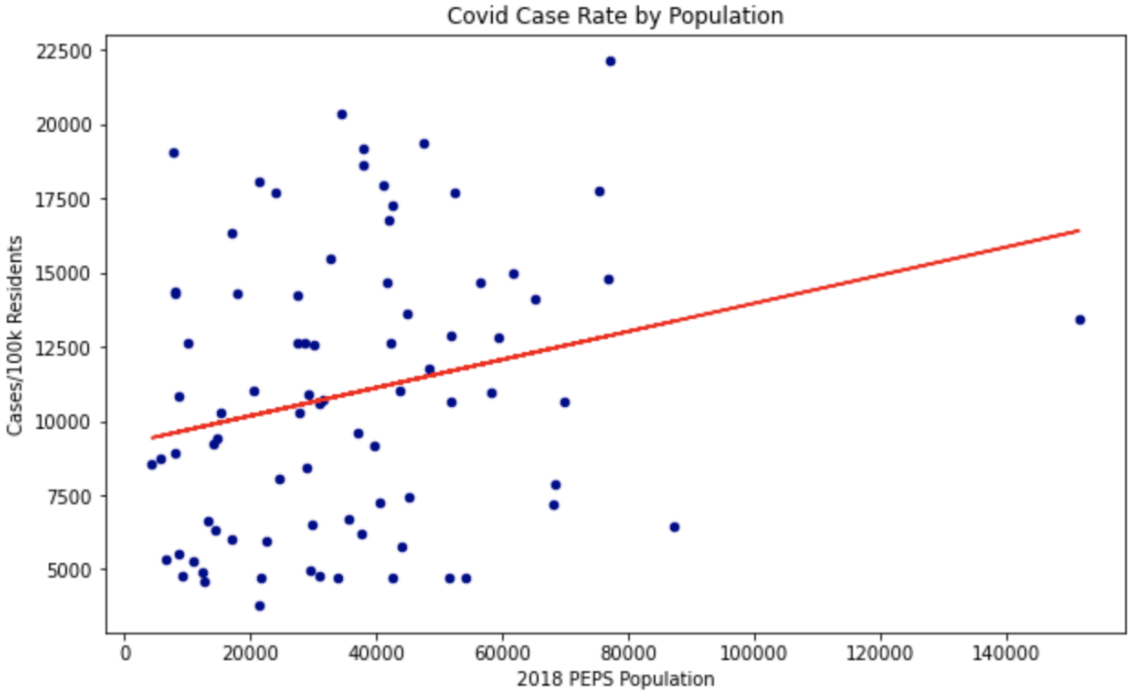
Fig. 1: COVID Case Rates by Neighborhood, Visualized



3.2 Population Density and COVID Rates

Another factor to consider was whether or not COVID rates were simply related to population density. I quickly plotted a scatter plot of COVID cases per 100k residents vs. population of a neighborhood, and performed a linear regression. See figure 2 for the result. You can see there was a slight positive correlation - this might be something to take into account in further analysis. However, with an R squared value of only 0.056, it seems unlikely this is the predominant factor in predicting COVID-19 cases, so we will dig in further.

Fig 2: COVID Cases per 100k Residents vs. Neighborhood Population



3.3 Round 1: K Means Clustering of Neighborhoods

Now that I examined the case data and how population density may affect COVID rates, I wanted to determine substantially similar and different clusters of neighborhoods by venue and how this affected COVID case rates. So, after gathering data on neighborhoods and venues and the FourSquare API, one-hot encoding this data, and taking the mean of venue types per neighborhood, I was ready to perform a first round of clustering using k-means.

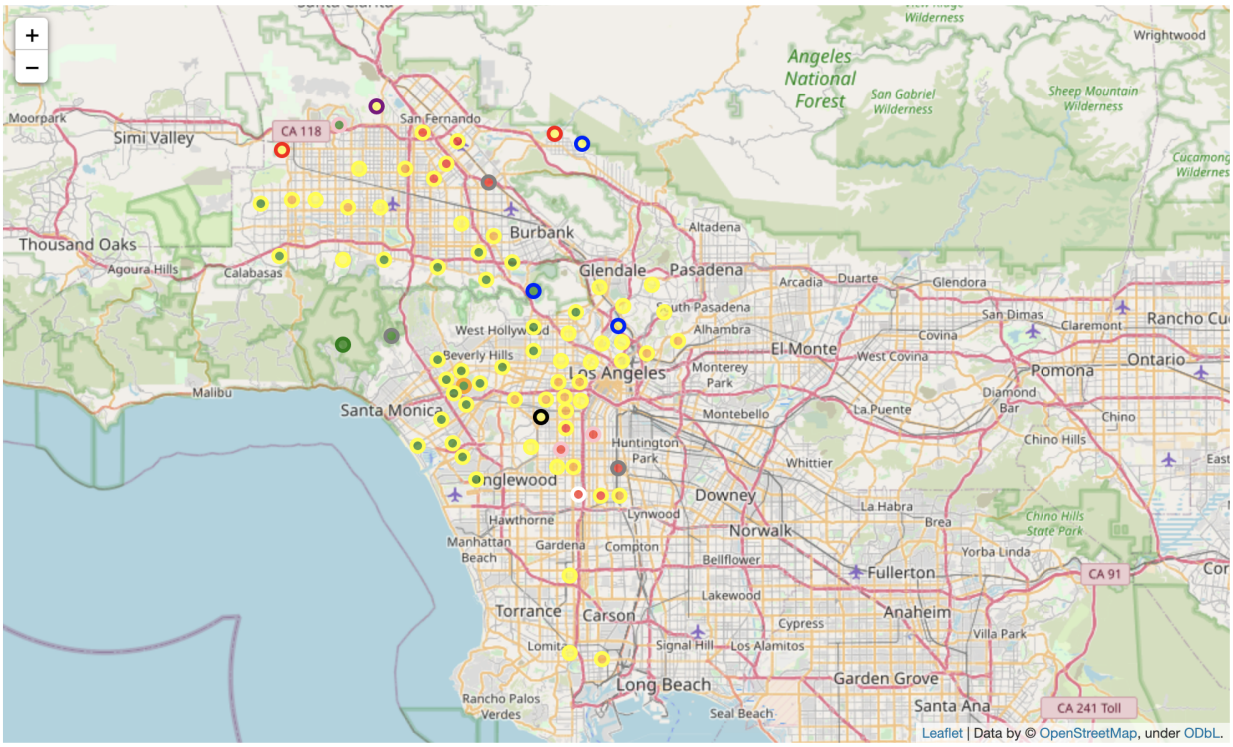
I used the k-means algorithm from the sklearn library with 10 clusters. I determined 10 when I realized I was essentially doing outlier detection - I tried 5-11 different clusters, and all but one cluster in each of these only contained 1 or 2 neighborhoods, almost exclusively on the outskirts of LA.

Viewing many of these individual points on a map, it becomes apparent that they are outliers. The majority (7 of 10 points) fall on the outskirts of LA proper, mostly in the hills. These represent either suburban or extremely affluent neighborhoods with (generally) lower case rates. See below to see how I determined this.

3.4 Conclusions from the first round of clustering

The points in South Park and Harvard Park (pink), Florence-Firestone and Sun Valley (grey), Leimert Park (black), and Vermont Vista (white) bear further discussion - they are deeply urban and have high COVID case rates. See figure 3 for a quick visualization of how neighborhoods were clustered by the first k-means algorithm on the map.

Fig 3: Results of the First K-Means Clustering



We can see from a quick analysis (and visually, at a glance) that clusters 4, 8, and 9 are the most heavily impacted by COVID.

Fig 4: Cases by Label, First K-Means Clustering.

	k-means-labels	Cases/100k Residents	label_color
0	8	17958.000000	white
1	4	14841.000000	pink
2	9	13964.333333	gray
3	5	10930.000000	purple
4	2	10909.682540	yellow
5	6	10305.000000	black
6	3	10283.500000	red
7	0	9292.666667	blue
8	7	4798.000000	orange
9	1	3790.000000	green

The types of venues in those locations can be seen by pulling all neighborhoods with those particular labels.

Fig. 5: Most Common Venues in Neighborhoods with Highest COVID Rates.

	k-means-labels	Neighborhoods	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
4	9	Brentwood	Food Truck	Scenic Lookout	Ethiopian Restaurant	Food Service	Food Court	Food & Drink Shop	Food	Flower Shop	Fire Station	Financial or Legal Service
20	9	Florence-Firestone	Food	Other Repair Shop	Music Venue	Grocery Store	Yoga Studio	Food Court	Food & Drink Shop	Flower Shop	Fire Station	Financial or Legal Service
57	9	Sun Valley	Food	Food Truck	Electronics Store	Taco Place	Fire Station	Furniture / Home Store	Seafood Restaurant	Donut Shop	Convenience Store	Film Studio

```
la_venue_types_by_neighborhood_sorted[la_venue_types_by_neighborhood_sorted["k-means-labels"]==8]
```

	k-means-labels	Neighborhoods	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
68	8	Vermont Vista	Burger Joint	Yoga Studio	Farmers Market	Food Stand	Food Service	Food Court	Food & Drink Shop	Food	Flower Shop	Fire Station

```
la_venue_types_by_neighborhood_sorted[la_venue_types_by_neighborhood_sorted["k-means-labels"]==4]
```

	k-means-labels	Neighborhoods	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
28	4	Harvard Park	Shipping Store	Park	Yoga Studio	Ethiopian Restaurant	Food Court	Food & Drink Shop	Food	Flower Shop	Fire Station	Financial or Legal Service
51	4	Porter Ranch	Park	Yoga Studio	Ethiopian Restaurant	Food Service	Food Court	Food & Drink Shop	Food	Flower Shop	Fire Station	Financial or Legal Service
55	4	South Park	Park	Yoga Studio	Ethiopian Restaurant	Food Service	Food Court	Food & Drink Shop	Food	Flower Shop	Fire Station	Financial or Legal Service

We can see from those clusters that several types of venues are indicators of high COVID rates. For this first round of clustering at least, a high concentration of Parks, Yoga Studios, and generic Food locations or Food courts indicates a higher incidence of COVID.

Aside from that, we've essentially detected 2 things with this clustering: 1) The boundaries of LA where the character of neighborhoods change dramatically; and 2) The affluent, diffuse neighborhood of Cheviot Hills.

In some ways, this is also useful info - all of the outliers not mentioned in the first paragraph are in the lower 2 tiers of COVID prevalence, telling me that suburban areas of LA have lower COVID rates.

We've already gleaned some interesting insights from our first round of clustering. However, most of LA is still in the "yellow cluster" - label 2. In our results, I do a deeper dive into all the points in yellow - that's likely where the meat of the problem is, and where we can truly figure out how composition of an urban neighborhood affects COVID outcomes.

5. Results

Figure 6 shows the results of the second round of clustering. There are now a few clusters in urban areas that line up with a higher incidence of COVID (keep in mind the fill color correlates with high or low incidence of COVID, where red has many cases and green has relatively few). One is the cluster of points outlined in red, and the other is the single point outlined in yellow. These clusters have a few things in common, which I will get into briefly. First, though, let's get some data on the clusters with the most cases and the least by determining the mean incidence of COVID by cluster and plotting it.

Figure 6: Results from the second round of clustering

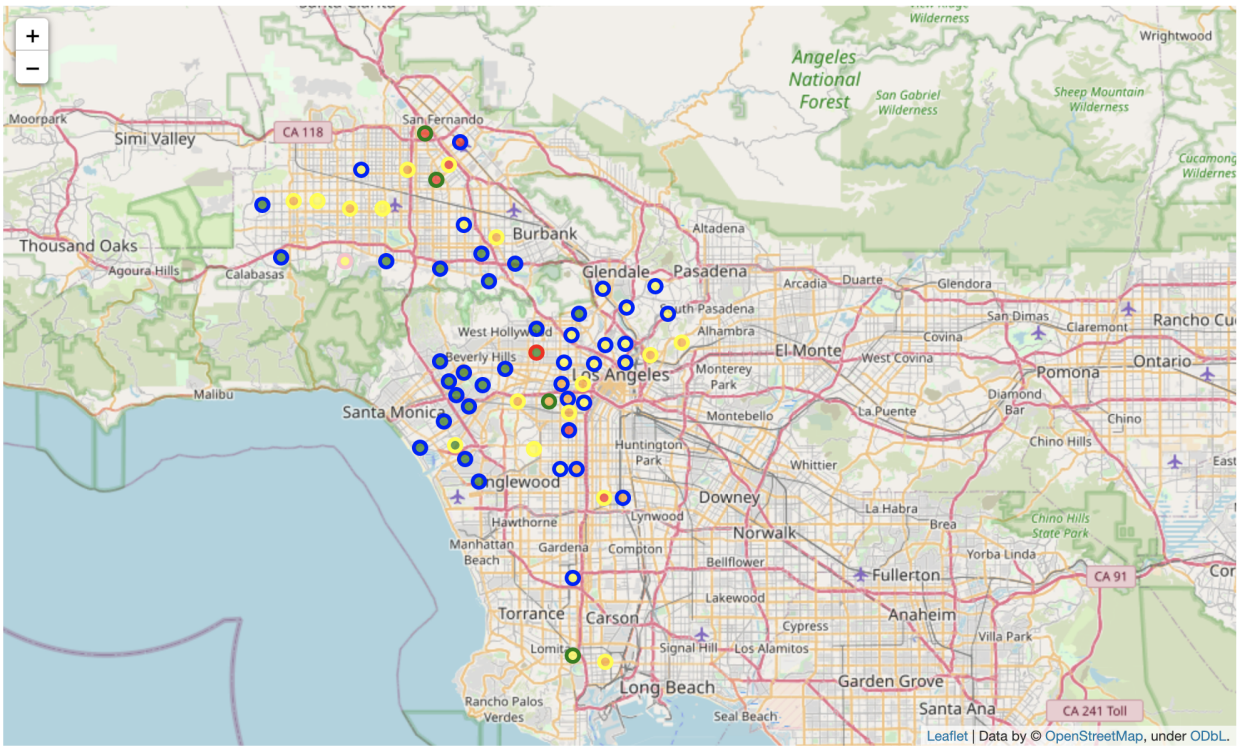


Figure 7: Clusters Ranked by COVID Prevalence

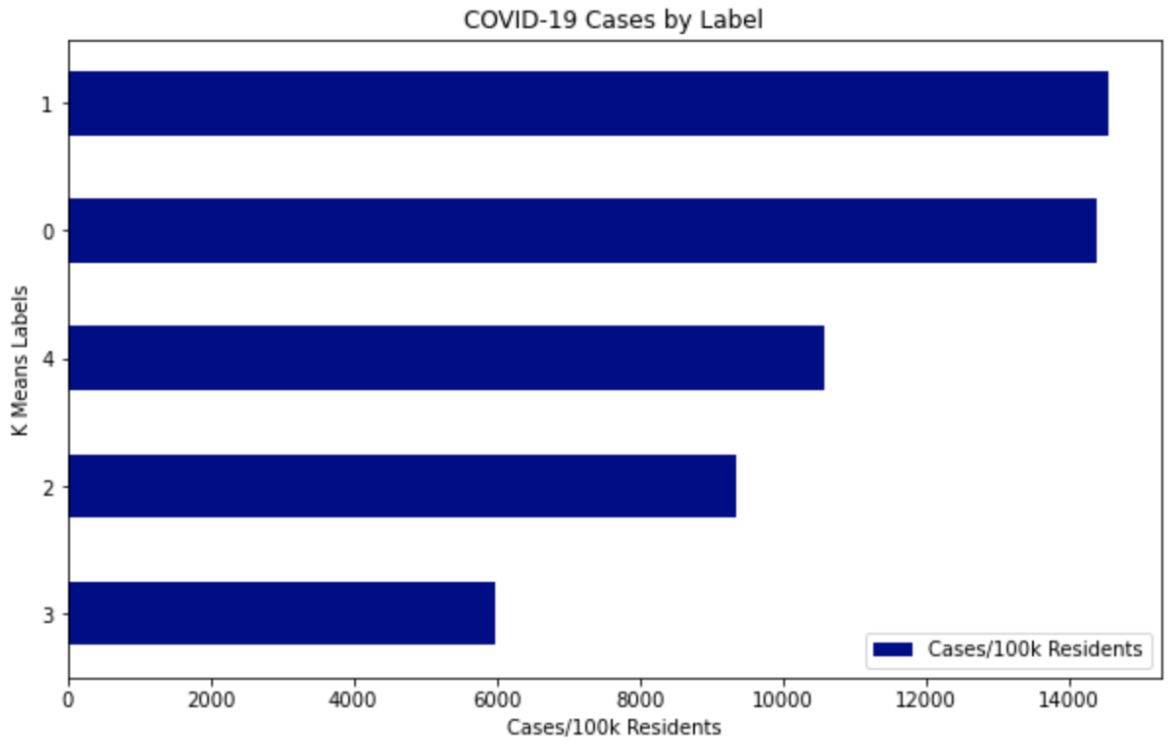


Fig. 8: Clusters Ranked by Mean Number of COVID Cases Per 100k Residents (color key in label_color column)

	yellow-k-means-labels	Cases/100k Residents	label_color
0	1	14542.250000	green
1	0	14367.000000	yellow
2	4	10568.000000	pink
3	2	9334.512195	blue
4	3	5986.000000	red

In figures 7 and 8, we can see that the clusters 0 and 1 have a high incidence of COVID compared to the others. Group 0 has an almost 36% higher incidence of COVID than the next group. By pulling out data for groups 0 and 1, we can, at a glance, determine the kinds of venues with high prevalence in those groups.

Fig. 9: Most Common Venues in Label 1

	yellow-k-means-labels	Neighborhoods	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
19	1	Harbor City	Mexican Restaurant	Bakery	Park	Spanish Restaurant	Yoga Studio	Farmers Market	Food Truck	Food Stand	Food Service	Food Court
25	1	Jefferson Park	Park	Neighborhood	Mexican Restaurant	Home Service	Taco Place	Fried Chicken Joint	Convenience Store	Yoga Studio	Fast Food Restaurant	Filipino Restaurant
32	1	Mission Hills	Plaza	Church	Park	Ethiopian Restaurant	Food Service	Food Court	Food & Drink Shop	Food	Flower Shop	Financial or Legal Service
38	1	Panorama City	Mexican Restaurant	Skating Rink	Curling Ice	Automotive Shop	Park	Yoga Studio	Falafel Restaurant	Food Service	Food Court	Food & Drink Shop

Fig. 10: Most Common Venues in Label 0

	yellow-k-means-labels	Neighborhoods	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
1	0	Arleta	Video Store	Historic Site	Bakery	Convenience Store	Farmers Market	Fast Food Restaurant	Filipino Restaurant	Financial or Legal Service	Yoga Studio	Ethiopian Restaurant
4	0	Canoga Park	Mexican Restaurant	Ice Cream Shop	Sports Bar	Restaurant	Furniture / Home Store	Sushi Restaurant	Liquor Store	Farmers Market	Food Service	Food Court
8	0	Del Rey	Mexican Restaurant	Hobby Shop	Bakery	Pizza Place	Donut Shop	Convenience Store	Coffee Shop	Sandwich Place	Café	Smoke Shop
12	0	El Sereno	Mexican Restaurant	ATM	Trail	Convenience Store	Restaurant	Neighborhood	Seafood Restaurant	South American Restaurant	Liquor Store	Thrift / Vintage Store
15	0	Exposition Park	Coffee Shop	Design Studio	Intersection	Mexican Restaurant	Food	Yoga Studio	Food Stand	Food Service	Food Court	Food & Drink Shop
17	0	Green Meadows	Food Stand	Pizza Place	Food	Donut Shop	Sandwich Place	Ethiopian Restaurant	Food Service	Food Court	Food & Drink Shop	Flower Shop
24	0	Hyde Park	Caribbean Restaurant	Convenience Store	Bookstore	Grocery Store	Yoga Studio	Falafel Restaurant	Food Stand	Food Service	Food Court	Food & Drink Shop
27	0	Lake Balboa	Sandwich Place	Steakhouse	Fast Food Restaurant	Convenience Store	Donut Shop	Automotive Shop	Seafood Restaurant	Mexican Restaurant	Flower Shop	Ethiopian Restaurant
28	0	Lincoln Heights	Mexican Restaurant	Fast Food Restaurant	Convenience Store	Burger Joint	Fried Chicken Joint	Music Venue	Gas Station	Sandwich Place	Food Truck	Video Store
33	0	North Hills	Pizza Place	River	Fast Food Restaurant	Baseball Field	Yoga Studio	Electronics Store	Food Service	Food Court	Food & Drink Shop	Food
34	0	North Hollywood	Latin American Restaurant	Shoe Store	Sandwich Place	Electronics Store	Mobile Phone Shop	Thrift / Vintage Store	Fast Food Restaurant	Donut Shop	Pizza Place	Insurance Office
39	0	Pico-Union	Latin American Restaurant	Mexican Restaurant	South American Restaurant	Cuban Restaurant	Clothing Store	Storage Facility	Convenience Store	Park	Grocery Store	Food
42	0	Reseda	Vietnamese Restaurant	Fast Food Restaurant	Mexican Restaurant	Furniture / Home Store	Chinese Restaurant	Supermarket	Convenience Store	Pawn Shop	Thai Restaurant	Falafel Restaurant
54	0	West Adams	Mexican Restaurant	Gym / Fitness Center	Fried Chicken Joint	Fast Food Restaurant	Latin American Restaurant	Bar	Performing Arts Venue	Café	Wine Bar	Deli / Bodega
60	0	Wilmington	Pizza Place	Fast Food Restaurant	Latin American Restaurant	Convenience Store	Discount Store	Mexican Restaurant	Museum	Park	Sandwich Place	Falafel Restaurant

Figures 9 and 10 show a few commonalities among neighborhoods with higher incidence of COVID. First, neighborhoods with a high incidence of COVID seem to have a high prevalence of Latin American and Mexican restaurants. More broadly, they tend to have many ethic

restaurants in general. The first, second, or third column almost always contains an ethnic restaurant for neighborhoods in these clusters.

The second is that, like in the first round of clustering, there are many outdoor venues appearing in the top 10 most common venues in these neighborhoods - parks, yoga studios, gyms, trails, etc. almost always show up early in the list of top 10 venues. These 2 points lead me to a few potential conclusions, which I summarize in the discussion and conclusion.

Looking at the remaining groups (feel free to see the notebook for more details), the first venues in a given neighborhood will almost never have a Mexican, Latin, or ethnic restaurant. The largest cluster in the second k-means grouping (2), has a high prevalence of Cafes and Coffee shops, banks and ATMs, and Theaters and Movie Theaters in the first through third most common venue. In the remaining 2 groups, each containing only one neighborhood, the most common venues were beauty salons and wine bars.

6. Discussion

We can see both in the table in Figure 8 and the graph in Figure 7 (COVID-19 Cases by Label) that labels 0 and 1 have by a significant margin (35% at least) the highest prevalence of COVID per 100k residents. Looking at the data in the tables in figures 7 and 8, we can conclude that urban LA neighborhoods with a high prevalence of ethnic restaurants (and in particular Mexican and Latin restaurants) have significantly higher prevalence of COVID-19 than other neighborhoods in LA. Case rates per 100k residents in these neighborhoods are higher than the rest of LA neighborhoods by more than a third.

From our first round of clustering, we can also see that locations with a high prevalence of Parks, Yoga Studios, and generic Food locations or Food courts indicates a higher incidence of COVID.

There are 4 hypotheses we might want to dig into to learn more.

1. This could be due to the fact that these neighborhoods had more traffic due to lockdowns and people only being allowed to engage in outdoor activities. This may have significantly increased people in those neighborhoods' likelihood of catching COVID.
2. This may be due to the demographics of these neighborhoods that correlate with these particular venues. Further investigation would be necessary to determine if income, age distribution, or other demographic factors played a role.
3. People from dense urban areas may have fled to these locations to be able to be outside during the pandemic.
4. These areas could have been less affected by lockdown, and more people went about their daily lives.

7. Conclusion

In conclusion, businesses looking to lower their risk of COVID going forward should choose to open in areas with fewer outdoor venues such as parks, trails, athletic courts, and historic sites.

This could either be because people brought COVID from other places to these outdoor areas as they tried to escape lockdowns, or may indicate that the demographics of these areas are such that they're more prone to the spread of COVID. For government agencies looking to dig further into these trends, neighborhoods in these clusters with high per capita COVID rates may be a good place to start looking for patterns of spread and reasons for increased spread.

We can also conclude that urban LA neighborhoods with a high prevalence of ethnic restaurants (and in particular Mexican and Latin restaurants) have significantly higher prevalence of COVID-19 than other neighborhoods in LA. Case rates per 100k residents in these neighborhoods are higher than the rest of LA neighborhoods by more than a third.

It's difficult to determine whether the cause of this distribution is first- or second- order. The trends I'm observing may be correlation. A deeper dive into demographics of these areas may be needed to truly determine the cause of this distribution. It may be worthwhile for municipalities to examine demographics in these areas to determine why they were so unequally affected by COVID.

8. Future Directions

There are many discoveries in this report that bear further investigation. It's likely worth digging into what it was that made the first k-means clustering we ran merely detect outliers, for instance, and what those outliers represent. We might also want to use demographic data from niche (https://www.niche.com/places-to-live/search/best-places-to-live/) to dig into demographic data for each cluster we discovered.

Some avenues worth exploring might be household income, median age, common ethnicities in given neighborhoods. These data would allow me to see if certain demographic groups were unevenly affected by COVID. I may investigate whether people in LA typically have primary residences close to their places of business to determine if the businesses in these locations are owned by and frequented by locals. I also might want to dig into how groups of different ethnicities fared during COVID in LA, and whether or not businesses closed in these locations due to COVID.

It is also worth mentioning that there is a weak relationship between population and COVID rates per capita. It might be worth digging into this independently as well - perhaps hardest hit areas were most affected by population density, and neighborhood composition was a secondary factor.

Finally, we may want to dig in to determine how people migrated during COVID, and if an influx of new people to given neighborhoods may have caused COVID cases to rise relative to the surrounding areas.