COVID-19 Spread Rates to Large Cities Popular Venues Correlation Study

Summary

Much has been documented about preventative measures to help individuals from catching the COVID-19 Virus, those range from hand washing, staying 6 feet apart, avoiding crowded places, isolating yourself, and wearing a mask.

January 21st, 2020 was the day the first case of COVID-19 was confirmed in the United States. The entire World has been affected by the virus, no place on earth is now free of it. Some communities have been affected more than others and researchers try to make sense of what affects certain cities more than others.

It is easy to assume the higher the population, the higher the rate of COVID-19 cases, but it does not end there. There are several studies, research papers, and articles studying population characteristics that contribute to the faster spread of COVID-19. Many agree that one of the reasons for some cities' infection rate larger than others cannot not be merely reduced to population size, or the density. The size of the family household has to be taken into consideration as well.

This study aims to not only analyze the relationship that population and average household size have on Coronavirus cases, but also checks if other population characteristics have an influential relationship. A different approach will be taken in this study. Not only demographics will be used in the comparison of cities, but also Foursquare API to explore the cities' most popular venues, and how that relates to the infection rate.

As you will see throughout the study, the type of the business can lead to a larger impact on the number of cases reported.

Interest:

This study is not trying to establish health recommendations; rather to provide a different perspective to individuals in our local and federal government involved in policy writing with respect to COVID-19. Additionally, it provides a different way of thinking about the problem to average citizens and researchers.

The Data:

Acquisition:

For this analysis, the data was acquired as follows:

- COVID-19 cases: Data was selected based on top 10 states affected by Coronavirus. With the CDC Covid Data Tracker, I was able to find links to each state's Coronavirus website. The data was at county level.

- County and City information: To find information about each county, I used Census Gazetteer Files. From there I was able to retrieve counties gazetter national files which had cities and counties as well as unique identifiers as FIPS and GEOID. I was able to also get places gazetteer files which provided me with cities latitude and longitude for geographic mapping.
- County and City Demographics: Demographics information, such as Population, Persons per Household, Density, and Median Income were collected from the census quickfacts page.
- Foursquare API was used to obtain the top 5 most common venues per city.

Methodology:

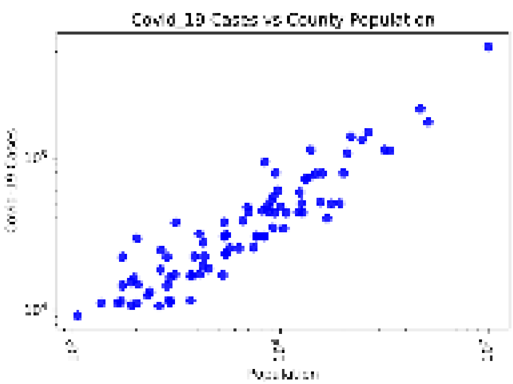
COVID-19 cases reported to the CDC were for the most part at county level, while our top venues from Foursquare API are at a city level. At this point I had city demographics information, and COVID-19 cases per county.

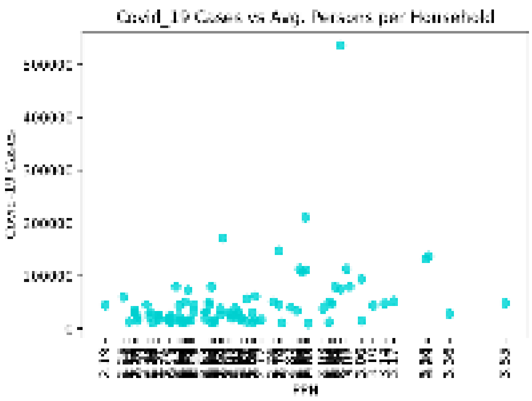
After researching and reading CDC guidelines on how COVID-19 cases spread, I selected variables that I thought seemed to have an impact on the increase of cases. The variables are: *Population estimates, Median Income, Density, and Persons per Household.*

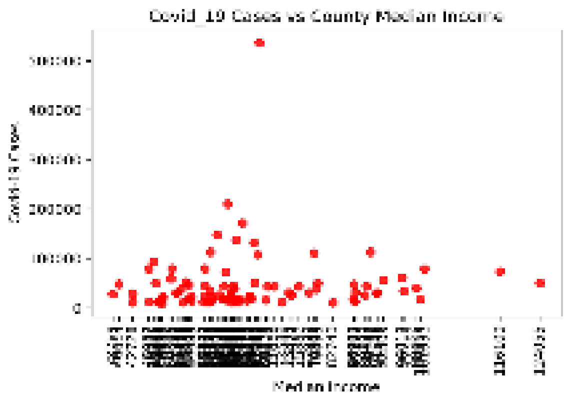
For this study, after merging county_df (dataset created after acquiring information from US Census Quickfacts) and cities_cty_df (a dataset created with the top 10 counties with highest COVID-19 cases per top 10 states), I started with a dataset that consisted of 88 observations and 8 columns.

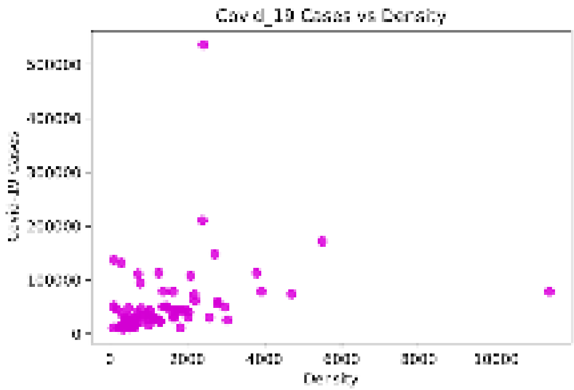
Once the dataset was completed, I started preparing it. This meant a transform of Population Estimate, Median Income, cases, Density, and Persons per Household from object to numeric. This data was at county level.

Relationship analysis was conducted for each of the features with reported COVID-19 cases, if any existed, and its significance. To find this possible correlation I created scatter plots to visualize the connection. Alsp, the Pearson Correlation Coefficient was compared as well as p-value to find any significance.

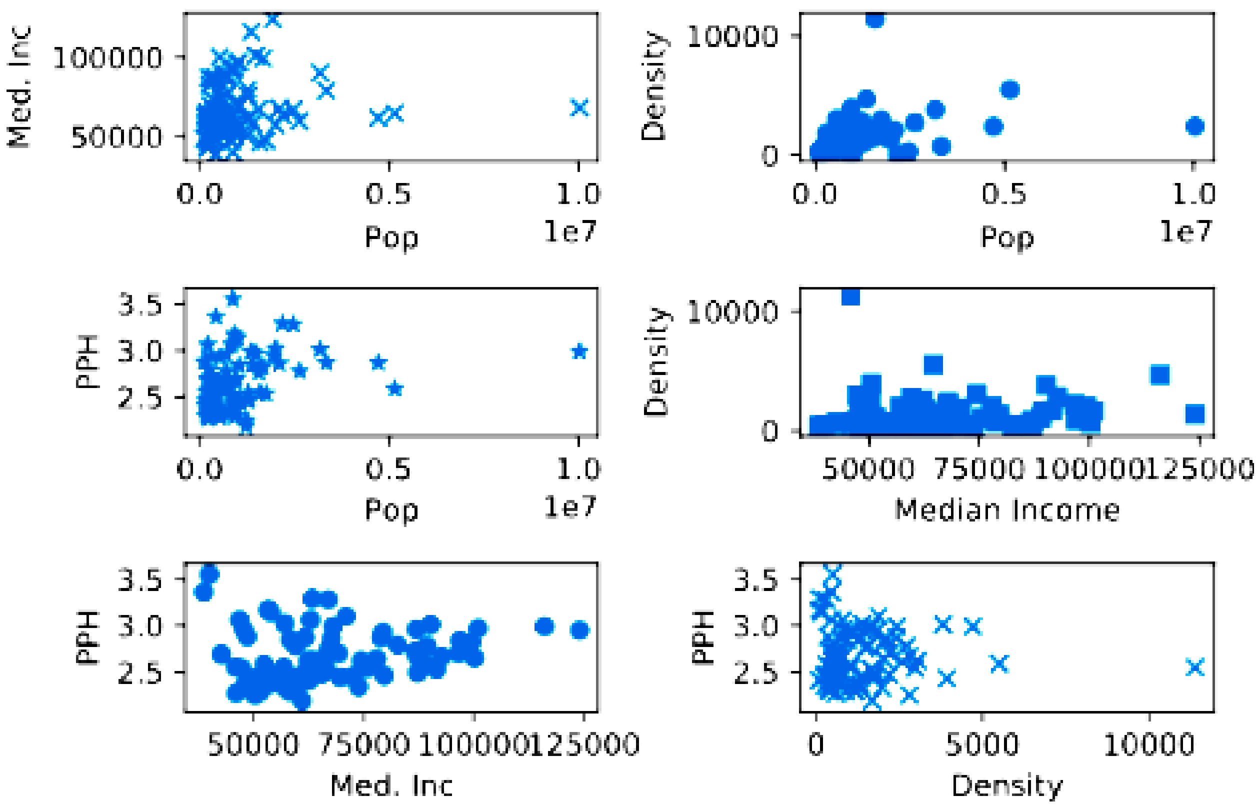


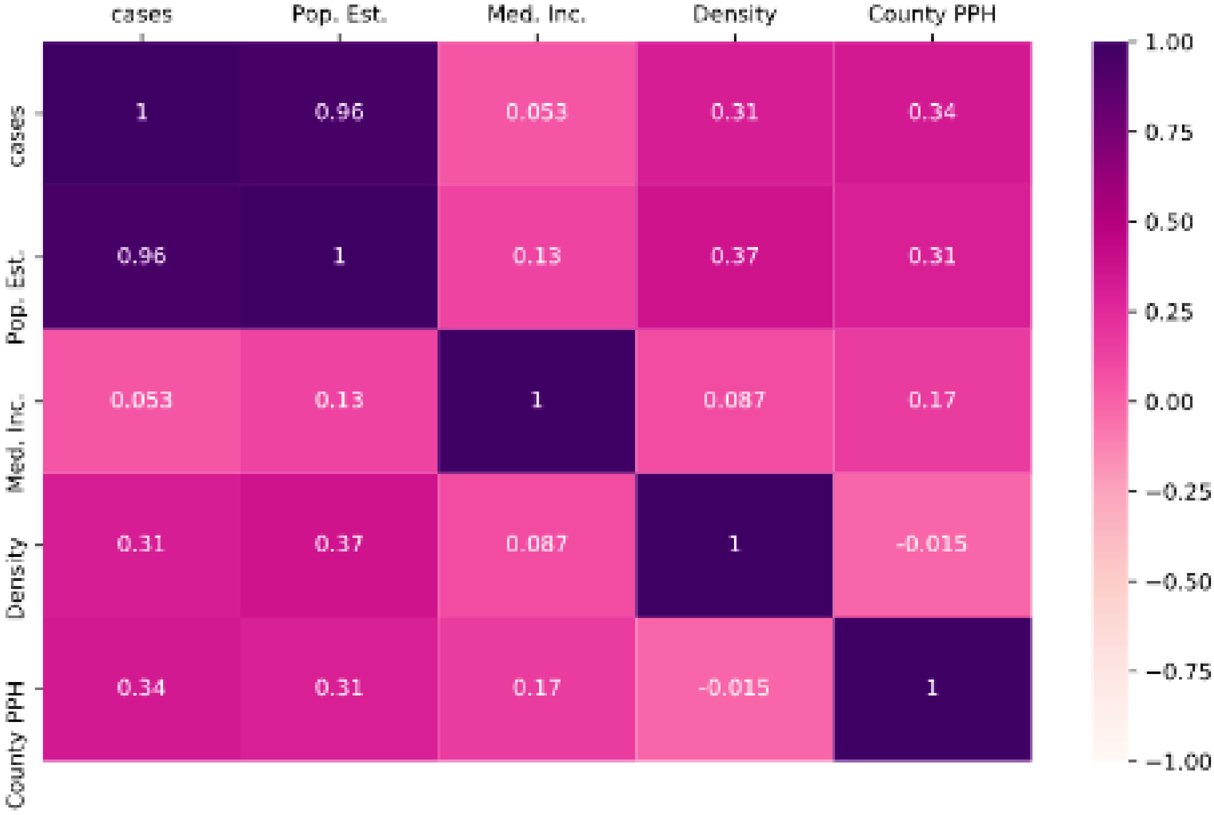






After all four features we ran individually against county cases variables, I ran them against each other to check for multicollinearity. At the end of the analysis, Population had the highest correlation with a Pearson Coefficient of 0.9579 and a p-value of 2.1856e-48 a second variable with a much lower coefficient is Persons per household with a coefficient of. 0.33 and a p-value of 0.0013. I also checked variables for VIF, variance inflation factor.





P-Values:

Pop_COVID-19 p-values = 2.185556571589477e-48

MedInc_COVID-19 p-values = 0.6237014516944078

PPH_COVID-19 p-values is: = 0.0013126784406875538

Density_COVID-19 p-values = 0.002926105376243104

	feature	VIF
0	Pop. Est.	1.901998
1	Density	2.062362
2	County PPH	2.077436

	cases	Pop. Est.	County PPH
cases	1.000000	0.957907	0.337256
Pop. Est.	0.957907	1.000000	0.310880
County PPH	0.337256	0.310880	1.000000

As we can see, population and PPH have the most significant p-values. We can also see that VIF values are low and no significant collinearity is found among these variables. Based on this I continue the study with Population Estimates and Persons per Household as my independent variables.

The focus at this stage is moved to city level. Here I started with three datasets:

- County COVID-19 cases (**county_covid19_data**)
- Cities in each county that I had on my county dataset (**places_df**)

- U.S Census City Population Estimates (**pop_est_df**)

After cleaning the last two datasets, and creating GEOID's for each city, a unique number to merge both datasets was used.

Combining county_covid19_data with places_df and then removing duplicate observations helped me drop the Census Designated Places as they are not legally incorporated and are created to provide data for settled concentrations of population. My dataset (named ccs_df) contained 2449 rows and 9 columns.

The ccs_df dataset was merged with population data. Next, I got a csv file with cities and their respective GEOID's, Latitude, Longitude values from the US Census website. This dataset was named **coords**. coords was merged with cities and population dataset that resulted in 1710 observations with 10 columns.

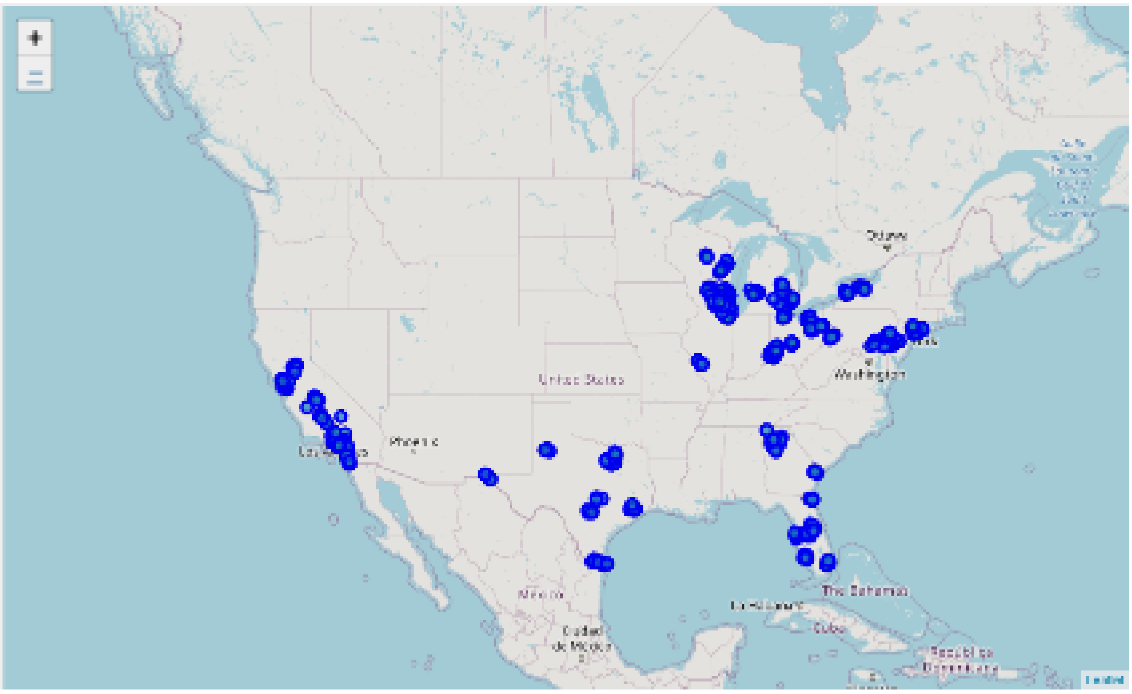
The dataset then was sorted in descending order by State, Cases, and Population. Being limited by Foursquare API to 950 calls quota per 24 hours, I had to reduce the number of observations to the top 6 rows per county reducing my dataset dimension to 513 x 10.

Persons per Household for each city is obtained from US Census quickfacts, and the help of request and regex library. I merged it with grouped_df1, giving me a total of 447 rows by 11 features/columns.

```
(447, 11)
```

	NAME	County	STNAME	STATE	TYPE	GEOID	POPESTIMATE2019	cases	LAT	LONG	Persons per Household
0	West Allis city	Milwaukee County	Wisconsin	WI	County Subdivision	5585300	59890	79398.0	43.007186	-88.026586	2.17
1	Wauwatosa city	Milwaukee County	Wisconsin	WI	County Subdivision	5584675	48118	79398.0	43.063165	-88.035583	2.33
2	Greenfield city	Milwaukee County	Wisconsin	WI	County Subdivision	5531175	37221	79398.0	42.950604	-88.005610	2.18

To help present the data in a visual formal, I created a United States Folium map that showed the cities.



Foursquare Analysis:

The Foursquare API was used to check for popular venues in each city with a radius of 500 meters. Cleaning, grouping, and sorting the data yielded a new dataset with City's name and the top 5 venues in that specific latitude and longitude with 389 rows and 6 columns.

	City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Ainsworth city	Beach	BBQ Joint	Park	Martial Arts School	Trail
1	Addison village	Business Service	Bank	Bar	Electronics Store	Gas Station
2	Akron city	Bar	Bank	Coffee Shop	Thai Restaurant	Sandwich Place
3	Allen city	Hotel	Breakfast Spot	Bowling Alley	Pharmacy	Zoo Exhibit
4	Allentown city	Racetrack	Business Service	Brewery	Eye Doctor	Fabric Shop

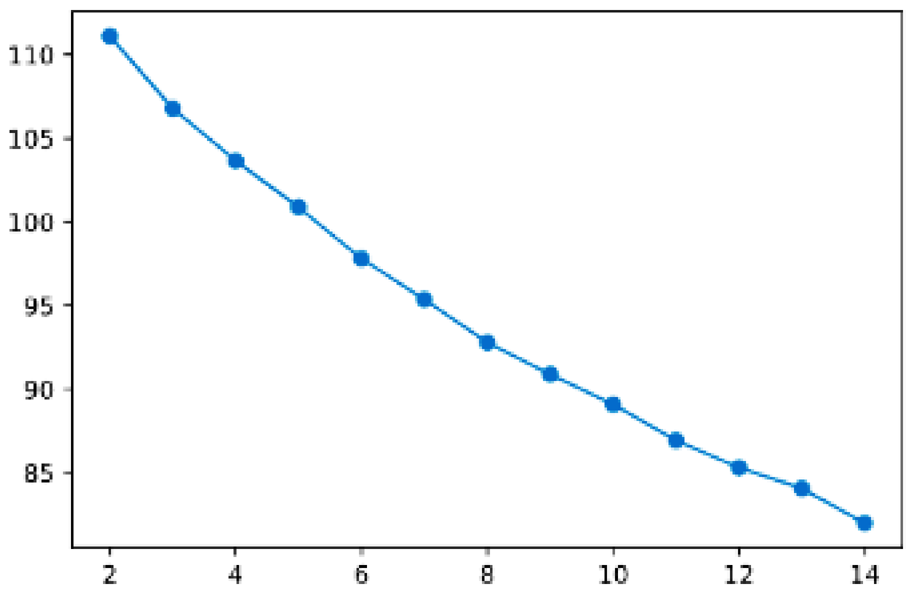
```
venues_sorted.shape
```

```
(389, 6)
```

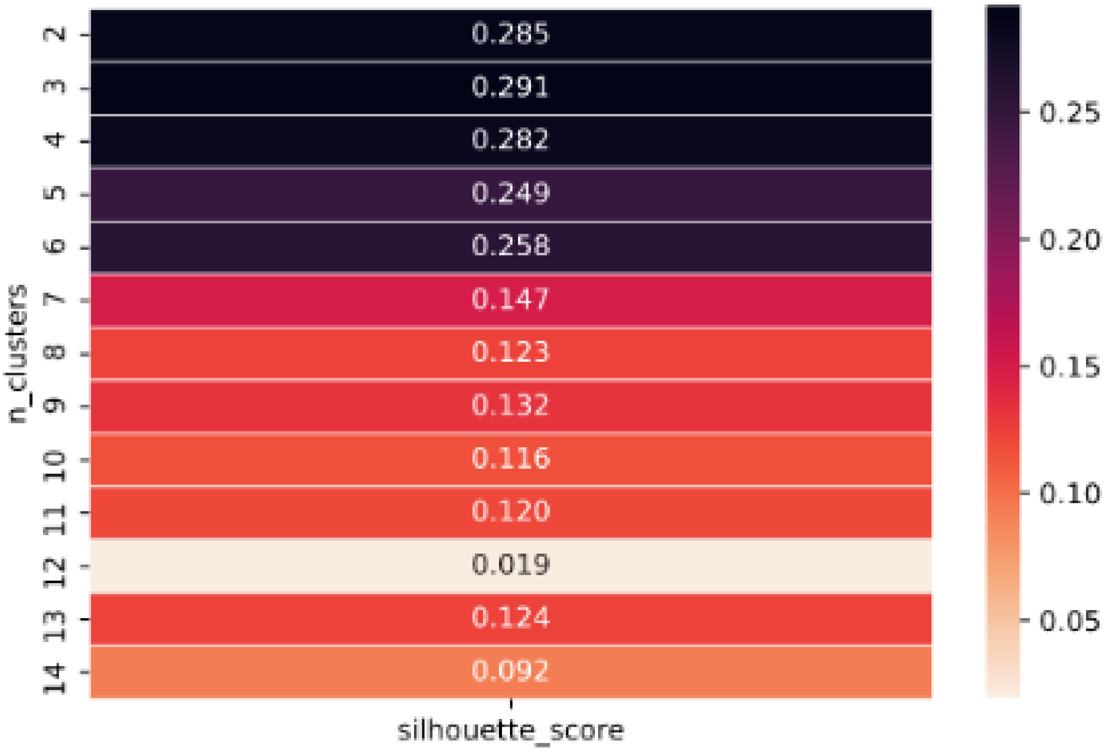
Clustering the Data

In order to find the optimal number of clusters I used both: The Elbow method and Silhouette Score.

Elbow Method produced the following plot:

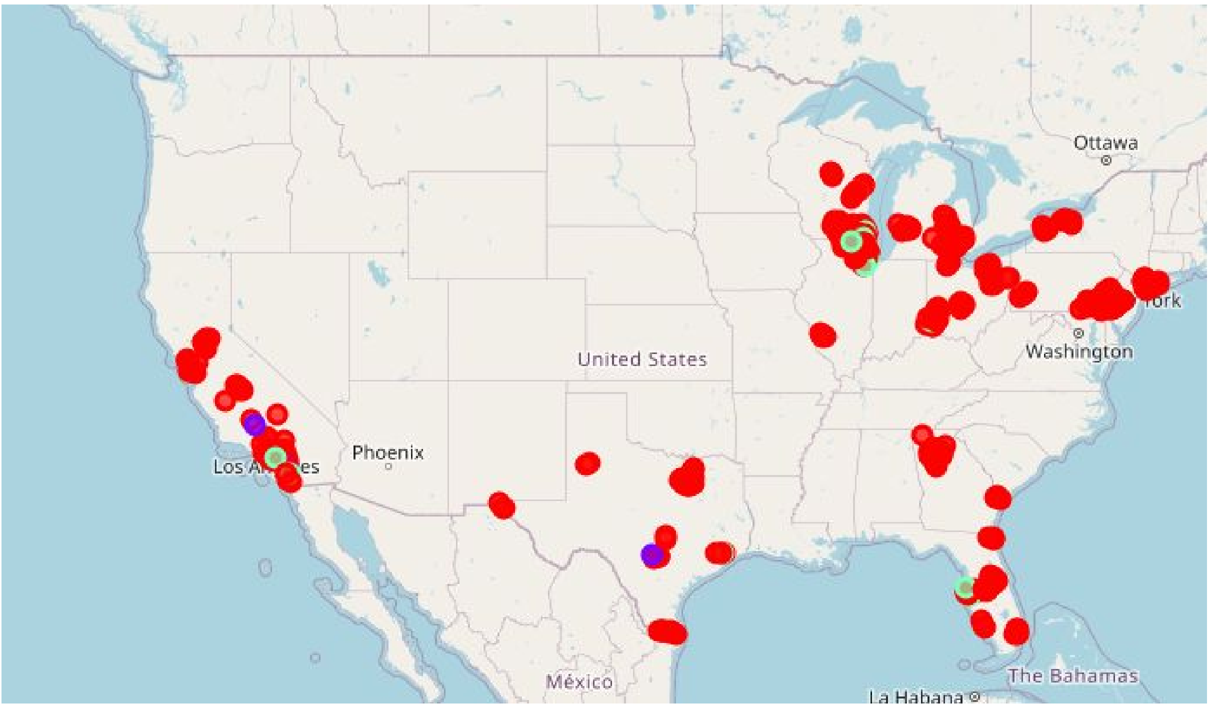


As you can see the decision is not very clear as to which point will be the optimal number of clusters. For this reason, I included a Silhouette score that produced the following table:



Based on Silhouette score, the optimal number will be at k_means cluster= 3. Then I added Cluster Labels to my data and continued to place the data on a folium map for visualization.

Clustering division is as follows:
Cluster_0 = 383
Cluster_1 = 4
Cluster_2 = 17



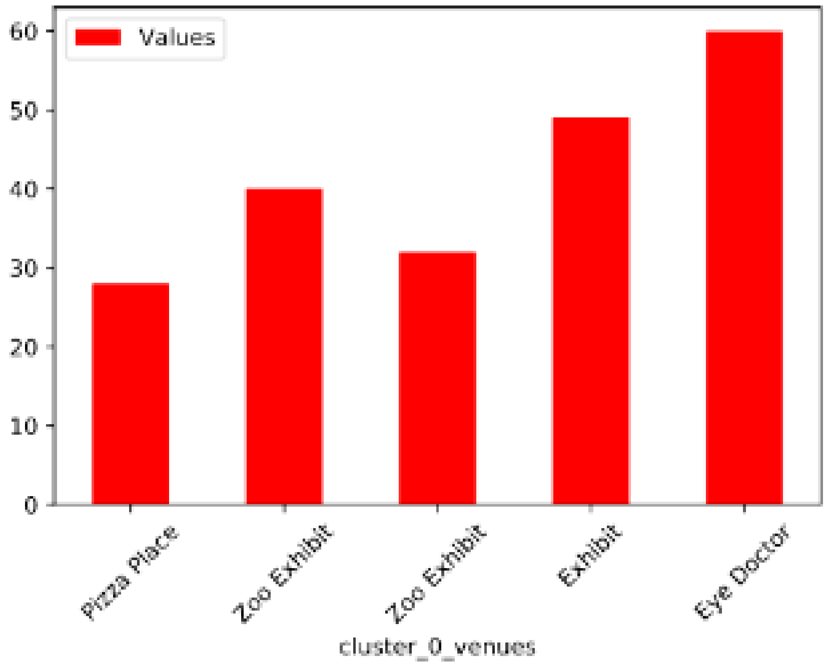
Findings

My focus was to find out if a pattern with popular venues and higher numbers of Population and Persons per Household can be visualised, which we know have an impact on COVID-19 cases.

Cluster_0

```
cluster_0[cols_list].describe()
```

	cases	POPESTIMATE2019	Persons per Household
count	383.000000	3.830000e+02	383.000000
mean	57226.704961	8.889624e+04	2.666632
std	71867.170852	2.525666e+05	0.401376
min	9165.000000	5.080000e+03	1.970000
25%	23394.000000	1.437200e+04	2.380000
50%	40026.000000	3.052800e+04	2.570000
75%	61111.000000	7.883450e+04	2.885000
max	536258.000000	3.979576e+06	4.280000

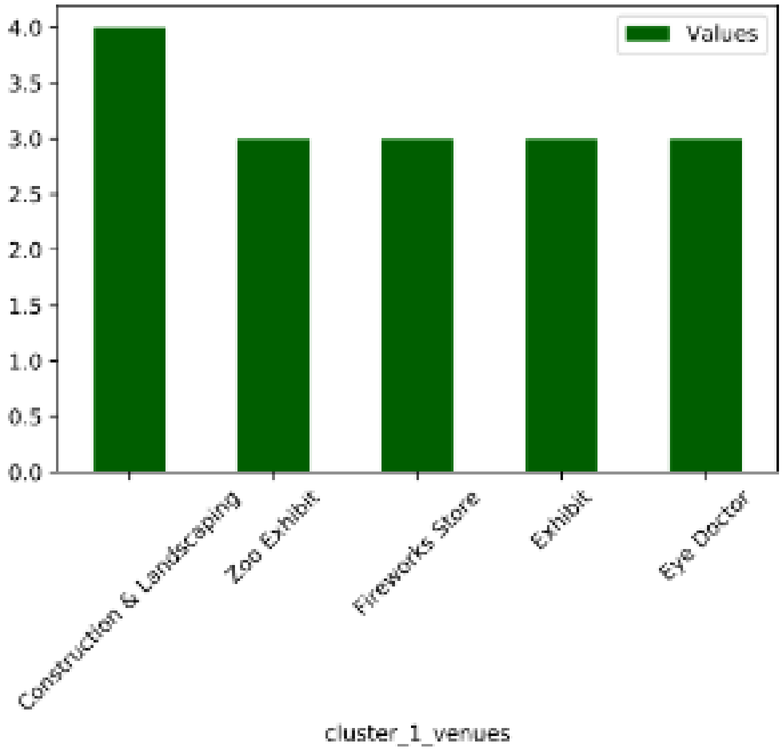


First cluster has an average Persons per Household value, and a high population. But cases are not as high when compared to the other two clusters. Though the number of observations probably plays a role in bringing down the case number values, I noticed the top venues are places that have been put under strict guidelines. Places like Zoos, Clinics were closed for a longer time at the beginning of the quarantine, then Pizza places are more of a pick-up or delivery business model, which helps keeping crowds small.

Cluster_1

```
cluster_1[cols_list].describe()
```

	cases	POPESTIMATE2019	Persons per Household
count	4.000000	4.000000	4.000000
mean	77511.750000	65879.750000	3.367500
std	38915.753043	57915.386323	0.805041
min	46860.000000	9961.000000	2.370000
25%	50716.500000	18878.500000	3.097500
50%	65682.500000	68306.000000	3.380000
75%	92477.750000	115307.250000	3.650000
max	131822.000000	116946.000000	4.340000

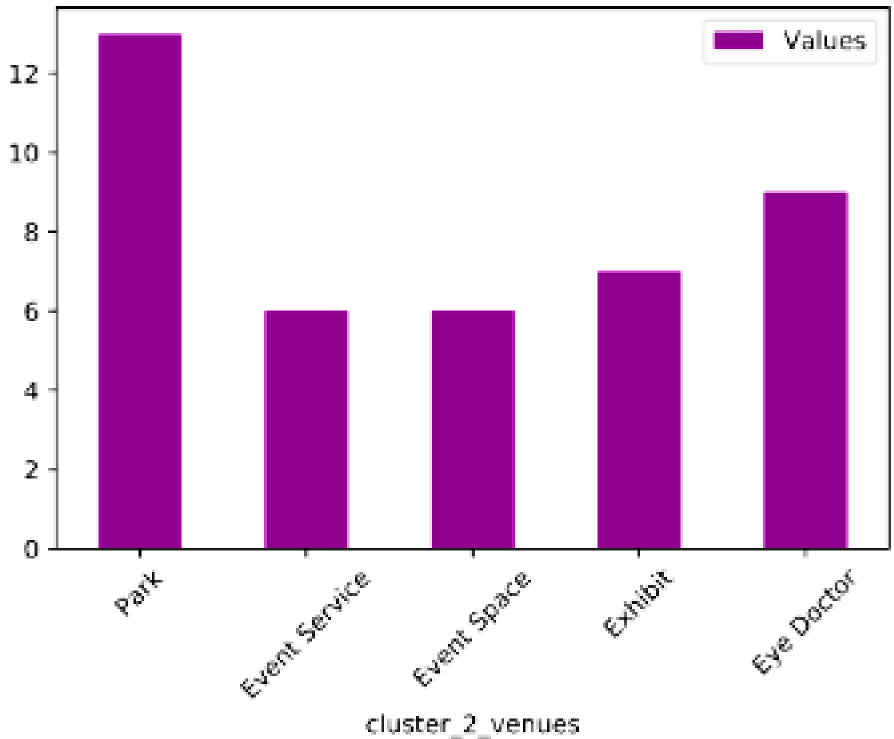


My second cluster has the highest average Persons per Household at 3.37, but it also is the smallest cluster with only 4 observations. Here the top venue/business is construction & landscaping. This particular business I have seen the staff go in groups, usually in the same vehicles. There is no remote work option for this type of business. Given the average size per household and the type of first venue, I believe COVID-19 cases spread faster.

Cluster_2

```
cluster_2[cols_list].describe()
```

	cases	POPESTIMATE2019	Persons per Household
count	17.000000	17.000000	17.000000
mean	68443.588235	60066.058824	2.658824
std	63521.130987	67605.807796	0.289501
min	11304.000000	5142.000000	2.300000
25%	17023.000000	20159.000000	2.440000
50%	46860.000000	34875.000000	2.530000
75%	111441.000000	62082.000000	2.870000
max	210362.000000	265351.000000	3.160000



The last group has the lowest Persons per Household average number and lowest population. Similar to the first cluster, the majority of popular venues are the kind of business that have been affected the most by quarantine shutdown. Parks are the exception, I have seen yellow tapes broken by people bringing their kids to play. Given that population and persons per household are the lowest in this cluster, I believe those two variables in combination with the type of top venues there would be a tendency of a slower COVID-19 cases spread in comparison with the other two clusters.

Discussion

Trying to find different virus propagation patterns that can help in the fight of COVID-19 cases is a worthy task. Understanding what aides in the propagation can help us determine better ways to slow it down.

This study is just a peak at what effect the type of most common businesses categories in a region can have on the number of cases, and it is a perspective that should be pursued. We see how a business model that requires people to be in the same vicinity can have an impact on the number of cases.

Some articles, like this by Bloomberg CityLab and this one by Business Insider, and studies have been done on how the number of people per household as well as population and density have an impact on the spread. These variables are ones that come to mind with more ease. The larger the pool of infected candidates the higher the number of cases. There are more variants that might have not been considered yet, and those are ones we need to include should this effort be continued by another researcher.

Listing some of the study limitations:

- For a more accurate picture, variables such as age, education, etc. should be considered.
- My API quota is limited. It may have been beneficial to have a larger sample.
- K-Means is relatively a simple model to use, but it has limitations when it comes to choosing the optimal number of clusters.
- Information on people working from home could have been useful.

Conclusion

There is no doubt the entire world is being affected by this pandemic, and the new variants are making the spread of the virus move at a faster pace. We have to follow guidelines and have common sense to have the least exposure possible. This study only looked into how the top businesses in an area could have an impact on the number of cases seen in an area.

The style of commercial activity in combination with the number of reported cases and other variables, could potentially show the impact of city policies and guidelines with respect to the spread of COVID-19.