Analysis of COVID-19 cases in Chicago Area

IBM Data Science Professional Certificate

Capstone Project

Ibrahim Benjelloun

5 November 2020

1. Introduction

The COVID-19 outbreak has been at the center of our preoccupations since the beginning of 2020. Up until now, more than 48M cases and 1.2M deaths have been registered worldwide. The United States of America has been severely hit as it is considered as the most impacted country in the world with more than 9 million infections and 230,000 deaths across the country.

Although important measures have been taken in the U.S. in an attempt to contain the outbreak, the country has reached a new record high in the number of daily new cases with 83,757 infections on October 23rd. The U.S. is on the verge of a third wave of the pandemic, and there are clear signs that this one is going to be worse than the first two waves.

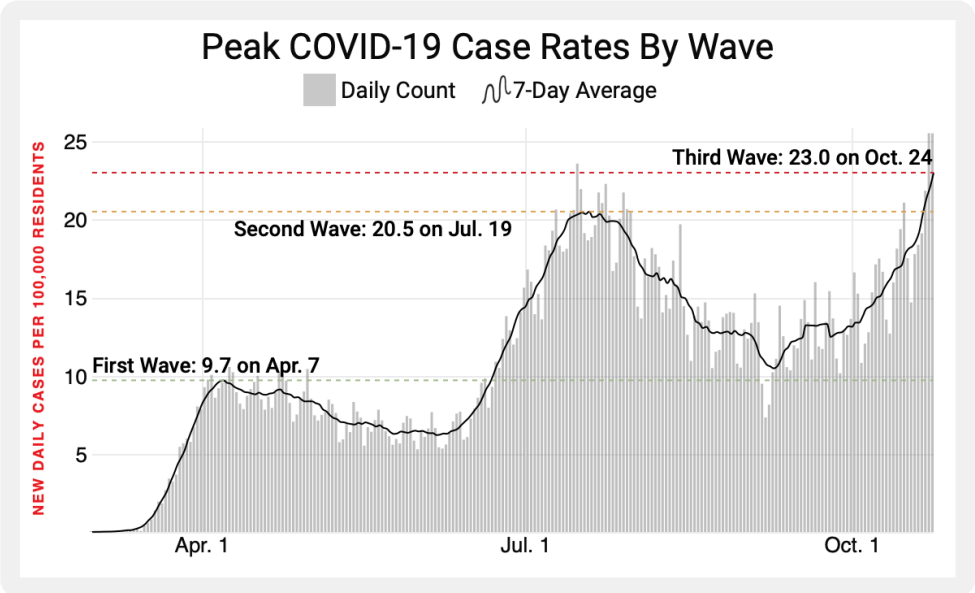

Fig 1: COVID-19 cases rates by wave in the U.S. [source]

In terms of number of cases, Texas comes first with approximately 950K cases followed closely by California (948K), Florida (816K), New York (513k) and finally Illinois (436K).

Various reasons could explain this continuous although wavy raise in daily COVID-19 cases. On the one hand, some say that it is largely due to behavioral reasons and the non-respect of Public Health measures such as social distancing and wearing face masks. On the other hand, some say it is also because of widespread testing and its increased efficiency.

Although lockdown measures have proven their efficiency in containing the outbreak, they are also not viable for many businesses that suffered economically and are still struggling to make up for the caused losses. This resulted in lifting lockdown restrictions and re-opening venues such as restaurants, bars, gyms, parks and others under the condition of putting in place sanitary measures to ensure the security and safety of the public.

The purpose of this project is to study an eventual relation between the concentration of certain types of venues in a specific area with the amount of COVID-19 cases registered in this area. This report is of course intended to anyone curious to know more about this subject, but specifically to local health authorities who want to better understand any correlation (if any) between the profile of an area based on the presence of certain types of “risky” venues and COVID cases.

The scope of this report is limited to the Chicago area. In the next chapters, we will first go through the data needed to complete the study. Then, we will describe the general methodology used to address this subject. Finally, we will discuss the results of the study and try to draw some conclusions.

2. Data

2.1. COVID-19 cases data for Chicago

First, we need some data about COVID-19 cases in the Chicago area. The Chicago Data Portal provides some interesting metrics about the outbreak in Chicago on a weakly basis from March 2020 until now. This dataset is interesting as it gives details by ZIP code. In fact, it is one of the main reasons why this study is performed on the city of Chicago as this level of granularity in COVID data is hard to find. Here are some fields examples:

- **ZIP code**: home ZIP code of the COVID cases and people tested.
- **Week number**: a sequential count of weeks, starting at the beginning of 2020.
- **Cases Weekly**: number of cases in the week
- **Cases Cumulative**: total number of cases through the weeks
- **Case Rate Weekly**: case rate per 100,000 population in the week
- **Case Rate Cumulative**: case rate per 100,000 population through the weeks
- **Population**: total population within a ZIP code area
- **ZIP code location**: geographic coordinates for the ZIP code area

For more information about the data and other fields, please refer to the official page.

	ZIP Code	Week Number	Week Start	Week End	Cases - Weekly	Cases - Cumulative	Case Rate - Weekly	Case Rate - Cumulative	Tests - Weekly	Tests - Cumulative	...	Test Rate - Cumulative
0	Unknown	10	03/01/2020	03/07/2020	NaN	NaN	NaN	NaN	5.0	5	...	0.0
1	Unknown	11	03/08/2020	03/14/2020	NaN	NaN	NaN	NaN	100.0	105	...	0.0
2	Unknown	12	03/15/2020	03/21/2020	NaN	NaN	NaN	NaN	333.0	438	...	0.0
3	60603	10	03/01/2020	03/07/2020	NaN	NaN	NaN	NaN	0.0	0	...	0.0
4	60601	24	06/07/2020	06/13/2020	6.0	77.0	41.0	524.7	105.0	781	...	5322.0
...	...	...	...	...	...	...	...	...	...	...	...	...

Fig. 2: COVID-19 Chicago cases after import

2.2. Venues data

In addition to COVID-19 cases per ZIP code, we need to retrieve nearby venues and related information such as their category, for each ZIP code area. This is possible through the Foursquare API. Using a developer account, it is possible to explore nearby venues given a specific location: category of the venue, name, likes, pictures, etc.

For the purpose of this study, for each Chicago ZIP code area, we will retrieve the names of the nearby venues, latitude, longitude and venue category.

	ZipCode	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	60601	41.886262	-87.622844	Chicago Architecture Center	41.887720	-87.623650	Arts & Entertainment
1	60601	41.886262	-87.622844	Harris Theatre for Music and Dance	41.883883	-87.621992	Arts & Entertainment
2	60601	41.886262	-87.622844	Chicago Cultural Center	41.883640	-87.624671	Arts & Entertainment
3	60601	41.886262	-87.622844	Chicago Architecture Biennial	41.884136	-87.624672	Arts & Entertainment
4	60601	41.886262	-87.622844	The Chicago Theatre	41.885539	-87.627151	Arts & Entertainment
5	60601	41.886262	-87.622844	Tiffany Dome At The Chicago Cultural Center	41.883481	-87.624693	Arts & Entertainment
6	60601	41.886262	-87.622844	Cloud Gate by Anish Kapoor	41.882668	-87.623319	Arts & Entertainment
7	60601	41.886262	-87.622844	McCormick Bridgehouse & Chicago River Museum	41.888858	-87.624777	Arts & Entertainment
8	60601	41.886262	-87.622844	Jay Pritzker Pavilion	41.882614	-87.621782	Arts & Entertainment
9	60601	41.886262	-87.622844	American Writers Museum	41.885640	-87.624673	Arts & Entertainment

Fig. 3: Venues data after import

There are 10 main venue categories in Foursquare, and each one has sub-categories:

- Arts & Entertainment (4d4b7104d754a06370d81259)
- College & University (4d4b7105d754a06372d81259)
- Event (4d4b7105d754a06373d81259)
- Food (4d4b7105d754a06374d81259)
- Nightlife Spot (4d4b7105d754a06376d81259)
- Outdoors & Recreation (4d4b7105d754a06377d81259)
- Professional & Other Places (4d4b7105d754a06375d81259)
- Residence (4e67e38e036454776db1fb3a)

- Shop & Service (4d4b7105d754a06378d81259)
- Travel & Transport (4d4b7105d754a06379d81259)

2.3. Chicago Choropleth map

For future needs in the study, we will display a Choropleth map of Chicago area. This requires external *".geojson"* file that was also downloaded from the Chicago Data Portal in the dedicated page.

3. Methodology

3.1. Data preparation

The first step in the study is cleaning the data and keeping only the information we need.

For the COVID-19 data, our analysis will be based mainly on the **Cumulative Case Rate per ZIP code**. Since we have no special interest in analyzing the cases by weeks, we removed all weekly indicators and grouped the lines by ZIP code to get the total case rate per ZIP code. Additionally, we extracted the longitude and latitude for each location. Here is an extract from the final COVID-19 data:

	ZipCode	CaseRateCumulative	Latitude	Longitude
0	60601	2194.2	41.886262	-87.622844
1	60602	2733.1	41.883136	-87.628309
2	60603	1533.2	41.880112	-87.625473
3	60604	4987.2	41.878153	-87.629029
4	60605	2238.5	41.867824	-87.623449

Fig. 4: COVID-19 data after processing

We also used the Foursquare API to retrieve the venues surrounding each ZIP code, their category and location (coordinates). For each ZIP code, we

retrieved the venues within a 500 meters radius. We retrieved a total of 5547 venues.

Once we stored the results in a data frame, we used the one-hot encoding technique to convert the categorical variables into a form that could be interpreted by a Machine Learning algorithm. The following table represents the resulting dataframe after applying the one-hot encoding technique on each venue for all ZIP codes, and grouping by ZIP codes to obtain the frequency of each venue category:

	ZipCode	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
0	60601	0.116732	0.116732	0.000000	0.116732	0.097276	0.116732	0.116732	0.085603	0.116732	0.116732
1	60602	0.115830	0.115830	0.000000	0.115830	0.115830	0.115830	0.115830	0.073359	0.115830	0.115830
2	60603	0.118577	0.118577	0.007905	0.118577	0.118577	0.118577	0.118577	0.043478	0.118577	0.118577
3	60604	0.116279	0.116279	0.007752	0.116279	0.116279	0.116279	0.116279	0.062016	0.116279	0.116279
4	60605	0.068085	0.127660	0.000000	0.127660	0.038298	0.127660	0.127660	0.127660	0.127660	0.127660
5	60606	0.046218	0.126050	0.004202	0.126050	0.126050	0.126050	0.126050	0.067227	0.126050	0.126050
6	60607	0.038462	0.164835	0.000000	0.137363	0.027473	0.109890	0.164835	0.071429	0.164835	0.120879
7	60608	0.074627	0.000000	0.000000	0.194030	0.029851	0.029851	0.238806	0.000000	0.343284	0.089552
8	60609	0.000000	0.000000	0.000000	0.235294	0.000000	0.000000	0.529412	0.000000	0.235294	0.000000
9	60610	0.020833	0.020833	0.000000	0.156250	0.067708	0.156250	0.156250	0.109375	0.156250	0.156250

Fig. 5: Frequency of venue categories per ZIP code

3.2. Machine Learning: K-Means clustering

Once our data was ready, we needed to understand how these ZIP code areas were similar in terms of venue categories. One way to do that is using **Unsupervised Machine Learning** and K-Means clustering to **create clusters based on the similarity between ZIP code areas**. K-Means algorithm is simple, efficient and perhaps the most commonly used technique for clustering.

However, K-Means can't be run without initially providing the numbers of desired clusters "K". Choosing the optimal "K" is very important as it directly influences the results of clustering. One can easily do that by computing the K-Means for a series values of "K" on the same dataset and compare the performances. The "Elbow method" and the "Silhouette method" are two

methods to evaluate the optimal “K", based respectively on the “Sum of Squared Errors” (SSE) and the “Silhouette Score”.

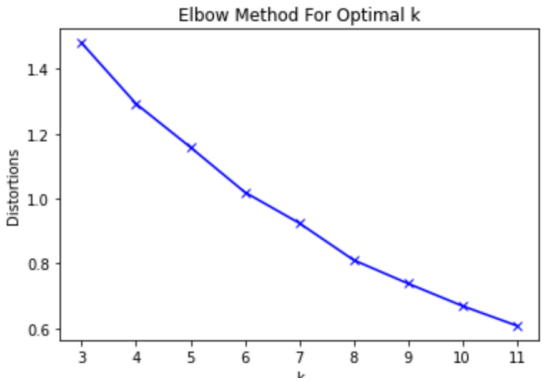

Fig. 6: Optimal K with the “Elbow method”

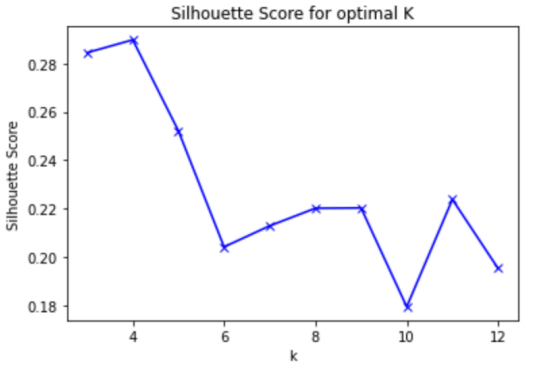

Fig. 7: Optimal K with the “Silhouette method”

In this case, we can barely distinguish the Elbow using the SSE whereas we see a clear pic of the Silhouette Score using **K=4 clusters**.

The preliminary results show the following characteristics for each cluster:

Cluster Label	Total Zip Codes	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	30	Shop & Service (**29.94%**)	Professional & Other Places (**26.44%**)	Food (**16.74%**)	Outdoors & Recreation (**8.21%**)	Travel & Transport (**5.45%**)
1	21	Shop & Service (**16.22%**)	Professional & Other Places (**17.78%**)	Food (**13.41%**)	Travel & Transport (**12.80%**)	Outdoors & Recreation (**10.15%**)
2	9	Professional & Other Places (**50.04%**)	Shop & Service (**16.22%**)	Food (**11.53%**)	Nightlife Spot (**4.72%**)	Outdoors & Recreation (**4.27%**)
3	1	Shop & Service (**60%**)	Food (**20%**)	Travel & Transport (**20%**)		

Fig. 8: Results of K-Means clustering

Focusing on the 3 most common venues, we notice that **the most common type of venues are “Shop & Service”, “Professional & Other Places” and “Food”**.

“Cluster 0” and “Cluster 1” have quite similar profiles in terms of top 5 venues. In “Cluster 0”, “Shop & Service”, “Professional & Other Places” and “Food” venues represent more than 70% of the venues whereas it represents less than 50% in “Cluster 1”.

"Cluster 2" has a very large number of "Professional & Other Places" venues that represent 50% of total venues, and "Cluster 3" distinguishes itself by the high proportion of "Shop & Service" venues.

The following map generated using the *Folium* library shows how the ZIP code ares are distributed within the four clusters:

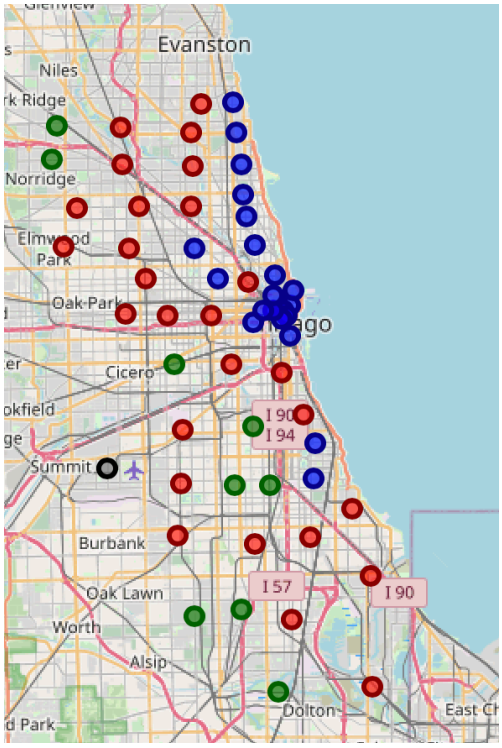

Fig. 8: Chicago's clusters. [Red: "Cluster 0", Blue: "Cluster 1", Green: "Cluster 2", Black: "Cluster 3"]

3.3. Chicago clusters and COVID-19 data

The first step in analyzing COVID data with the resulting clusters was appending the Cluster Labels to our data. Then we generated a Choropleth map to visualize how the COVID case rates change between each cluster's ZIP code areas:

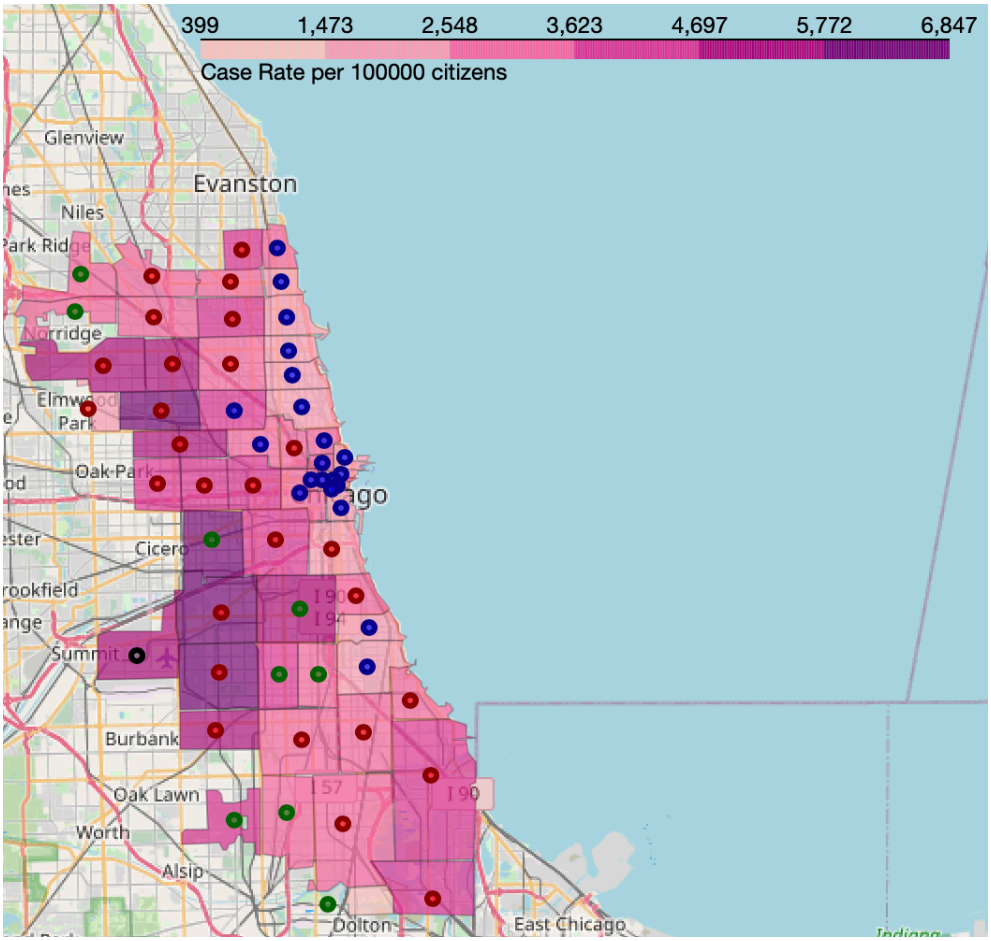

Fig. 9: Choropleth map of Chicago's clusters and case rate per 100,00 population per ZIP code. [Red: "Cluster 0", Blue: "Cluster 1", Green: "Cluster 2", Black: "Cluster 3"]

For further analysis, we also computed the average case rates per cluster as following:

Cluster Label	Total Zip Codes	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Average Case Rate
0	30	Shop & Service (**29.94%**)	Professional & Other Places (**26.44%**)	Food (**16.74%**)	Outdoors & Recreation (**8.21%**)	Travel & Transport (**5.45%**)	3970.592593
1	21	Shop & Service (**16.22%**)	Professional & Other Places (**17.78%**)	Food (**13.41%**)	Travel & Transport (**12.80%**)	Outdoors & Recreation (**10.15%**)	2788.690476
2	9	Professional & Other Places (**50.04%**)	Shop & Service (**16.22%**)	Food (**11.53%**)	Nightlife Spot (**4.72%**)	Outdoors & Recreation (**4.27%**)	3717.877778
3	1	Shop & Service (**60%**)	Food (**20%**)	Travel & Transport (**20%**)			5218.000000

Fig. 10: Results of K-Means clustering with average case rate per cluster

Then, we plotted the same Choroplteth map with the new average indicator:

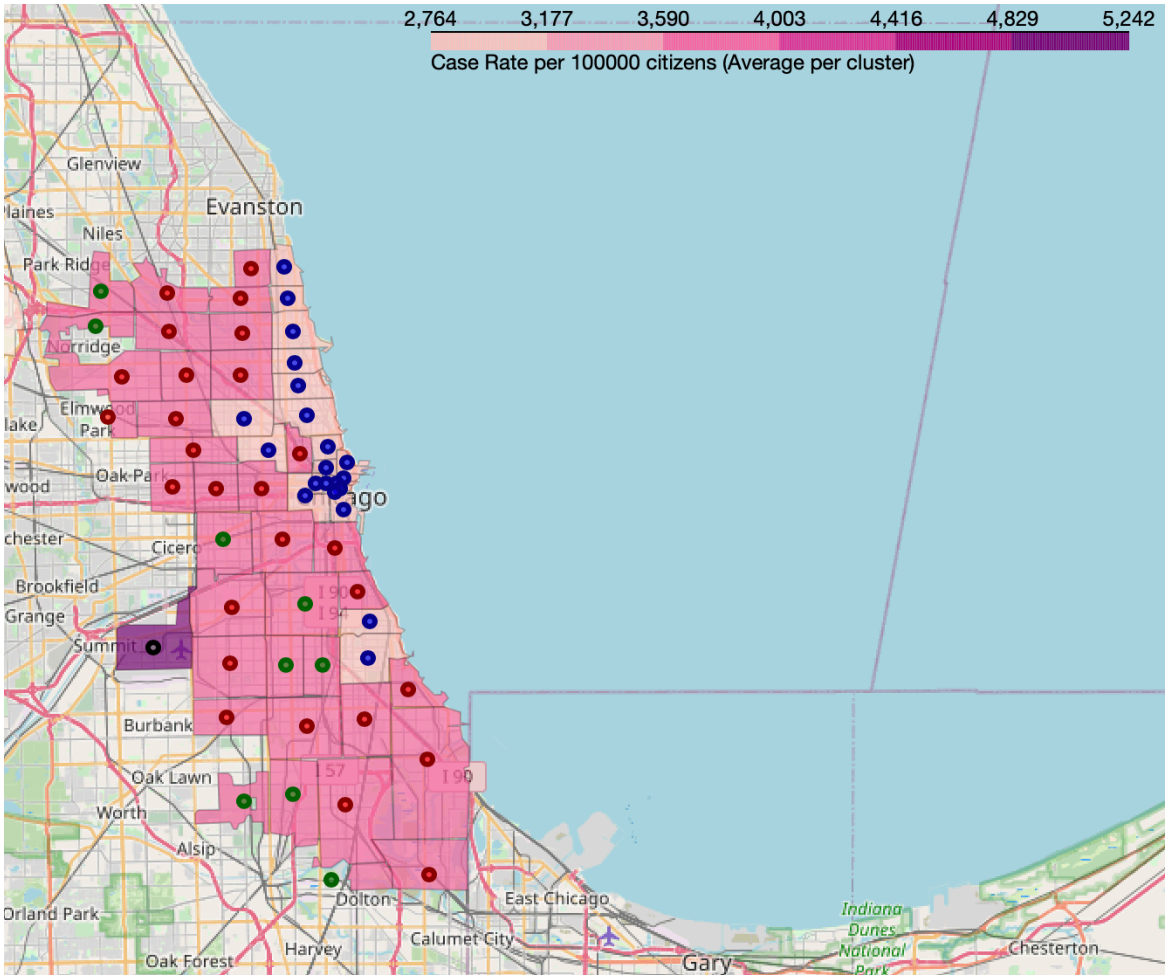

Fig. 11: Choropleth map of Chicago's clusters and average case rate per cluster. [Red: "Cluster 0", Blue: "Cluster 1", Green: "Cluster 2", Black: "Cluster 3"]

4. Results

- "Cluster 1" is the least impacted cluster with an average of 2788 cases per 100,000 population. When it comes to the top 5 most common venues, this cluster is the most homogenous one as the frequencies of each venue varies between 10% and 16%.
- Following is the "Cluster 2" with an average of 3717 case rate. The main difference with "Cluster 1" is the considerable proportion of "Professional & Other Places " venues with 50%.

- "Cluster 0" is slightly more impacted than "Cluster 2" with 3970 average case rate. In comparison with "Cluster 1", "Shop & Service" and "Professional & Other Places" venues proportions have shifted to nearly 30% each.
- Finally, "Cluster 3" is the most impacted one with 5218 average case rate. This could be interpreted with the very important proportion of "Shop & Service" venues (60%) and also a relatively large proportion of "Travel & Transport" venues.

5. Discussion

From these results and given the data we had, one could say that an area with homogenous proportions of venue types are generally less affected by COVID-19 as per **"Cluster 1"**. These kind of areas can be compared to a regular **Residential Area** with shops, offices, restaurants, etc.

One could also say that COVID-19 case rates are very sensitive to the proportion of "Professional & Other Places" and "Shop & Service" venues as per "Cluster 2" and "Cluster 0", and that these two types venues have almost similar impacts. **"Cluster 0"** could be assimilated to a **Commercial Area** and **"Cluster 2"** to a **Business Area.**

Finally, an area with high proportion of "Shop & Service" and "Travel & Transport" venues has proven to be highly impacted by COVID-19 as per **"Cluster 3"**. These areas can be assimilated to an **Airport Zone** where we tend to have high density of shops, hotels but also transportation services. ***Looking at the Choropleth map, one can confirm that "Cluster 3" is near the Chicago Midway International Airport***.

However, as much as Foursquare data gave us some useful directions for our study, one can say that it lacks some insightful informations such as the size of a venue, its daily traffic, etc. Also, although COVID data was some of

most detailed ones available to public that could be found, it remains obvious that a more interesting and complete approach would be to explore data at individual level for positive COVID cases, with the type of venues visited during a specified time frame before the testing (1 week for example). Combining these data would be much more insightful although more complex.

6. Conclusion

The results obtained in this study are high level and more qualitative than quantitative. However, they confirm some of our intuitions regarding the risk of getting infected depending on the type of area we are. An interesting thing to do would be to confirm these results by comparing the data of another area and see if our conclusions apply there.

From a personal perspective, this is my first Data Science project as part of the IBM Professional Certificate. It has been a very rich journey and amazing introduction to Data Science and Machine Learning.

For more information, please visit the Git repository of this project.