Choosing a place to live in Rio de Janeiro during the COVID-19 pandemic

Romeu Ribeiro Marques da Fonseca
Brasília, April 2nd, 2021

1. Introduction

During the COVID-19 pandemic, someone who needs to choose a neighborhood in the city of Rio de Janeiro to live, may question which areas are safe or not.

There are many aspects to consider when choosing a safe area in a city regarding the pandemic: the demographic density of the region (lower is better), the presence of commercial establishments and tourists spots - which attracts people and improves the chances of spreading the virus, the average age of residents and other risk groups, the medical infrastructure available, and many others.

In this final assignment, we are going to visualize each neighborhood of Rio de Janeiro by the COVID-19 death rate and the distribution of the city's medical infrastructure (hospitals, urgent care centers and emergency rooms) as the criteria to select the best regions to live.

This is just one criterion amongst many others that are related to death or recovery cases of infected people.

As this study progresses, we will discover that good medical infrastructure in certain areas can help, as expected, in the recovery of patients. On the other hand, the lack of proper medical care can directly influence on a larger death count.

2. Data

The geospatial data, as well the COVID cases data, are from official institutions of the local government. The datasets can be obtained in the following URLs:

- http://dadosabertos.rio.rj.gov.br/apiUrbanismo/apresentacao/csv/bairros_.csv
- https://www.data.rio/datasets/cep-dos-casos-confirmados-de-covid-19-no-munic%C3%ADpio-do-rio-de-janeiro
- https://www.data.rio/datasets/limite-de-bairros?geometry=-44.899%2C-23.138%2C-41.992%2C-22.695 (The geoson file with the neighborhood limits)

The first link has the CSV file with the coordinates of all Rio de Janeiro's neighborhoods, with its identification.

The second link has all the data of the COVID-19 cases in the city since the start of the pandemic. It is possible to isolate each case by the neighborhood where it was identified.

The last link has the geoson data where we can define the limits of each neighborhood. It will be important when plotting the Choropleth Map in the end of the study.

Since the data come all from the same source – the local government, we should expect that it is completely standardized. In fact, it is not. There are neighborhood misspellings and other minor corrections that were made accordingly.

In the end, after all data cleansing, the final dataframe achieved was called *dfRioData*. The death rate is a crucial information to the project, and it is determined by calculating the death cases divided by the total cases (the sum of dead and recovered patients).

After collecting all the COVID-19 and geospatial data, Foursquare API will be used to gather the medical infrastructure in the region – the presence of hospitals, urgent care centers and emergency rooms - so it will be possible to conduct the final study. The quantity of each medical infrastructure will be the features in the K-Means algorithm.

3. Methodology

Once the Foursquare data is collected, we will convert each category of medical infrastructure in numerical values using the one hot encoding technique.

All this information is stored in a dataframe: *rio_onehot*.

We will, then, use the K-Means algorithm to partition the neighborhoods in 3 groups: cluster 0, cluster 1 and cluster 2. The clustering process will be based on the quantity of Hospitals, Urgent Care Centers and Emergency Rooms nearby, using the *rio_onehot* dataframe as the features source.

From the cluster information and geospatial data, we are going to inspect all regions in Rio de Janeiro by COVID-19 death rate and its relationship with the clusters of neighborhoods with similar medical infrastructure.

Later, all this information will be analyzed in a Choropleth map where we can visually verify whether there is any correlation between the deaths by COVID-19 and the medical infra.

3.1. Exploratory Data Analysis

Before beginning the clustering process, we have built an initial Choropleth map showing which neighborhoods were most affected considering the pandemic. In this case, we are only considering the top 10 regions where more than 1000 cases were reported.

	ID	Neighborhood	Latitude	Longitude	Death	Recovered	Total Cases	Death Rate	Recovery Rate	Death Rate(%)
54	378	PADRE MIGUEL	-22.8833333	-43.45	207.0	950.0	1157.0	0.178911	0.821089	17.891098
159	511	BANGU	-22.883333	-43.466667	634.0	3517.0	4151.0	0.152734	0.847266	15.273428
123	450	GUARATIBA	-22.9984858	-43.5799151	215.0	1206.0	1421.0	0.151302	0.848698	15.130190
63	389	SENADOR CAMARA	-22.8666667	-43.5	224.0	1273.0	1497.0	0.149633	0.850367	14.963260
6	370	REALENGO	-22.8833333	-43.4333333	492.0	2906.0	3398.0	0.144791	0.855209	14.479105
158	510	CAMPO GRANDE	-22.8825	-43.5625	943.0	5716.0	6659.0	0.141613	0.858387	14.161285
29	356	GUADALUPE	-22.8372222	-43.3752778	140.0	959.0	1099.0	0.127389	0.872611	12.738854
82	406	COSMOS	-22.9166667	-43.6166667	129.0	927.0	1056.0	0.122159	0.877841	12.215909
31	351	BRAS DE PINA	-22.8316343	-43.293239299999996	150.0	1082.0	1232.0	0.121753	0.878247	12.175325
15	337	PAVUNA	-22.8136111	-43.3602778	164.0	1232.0	1396.0	0.117479	0.882521	11.747851

Figure 1 - Top ten regions with more than 1000 cases of COVID-19

From the table above we see that the top ten neighborhoods ordered by the death rate ranges from almost 12% to 18%.

The resultant Choropleth map is the following:

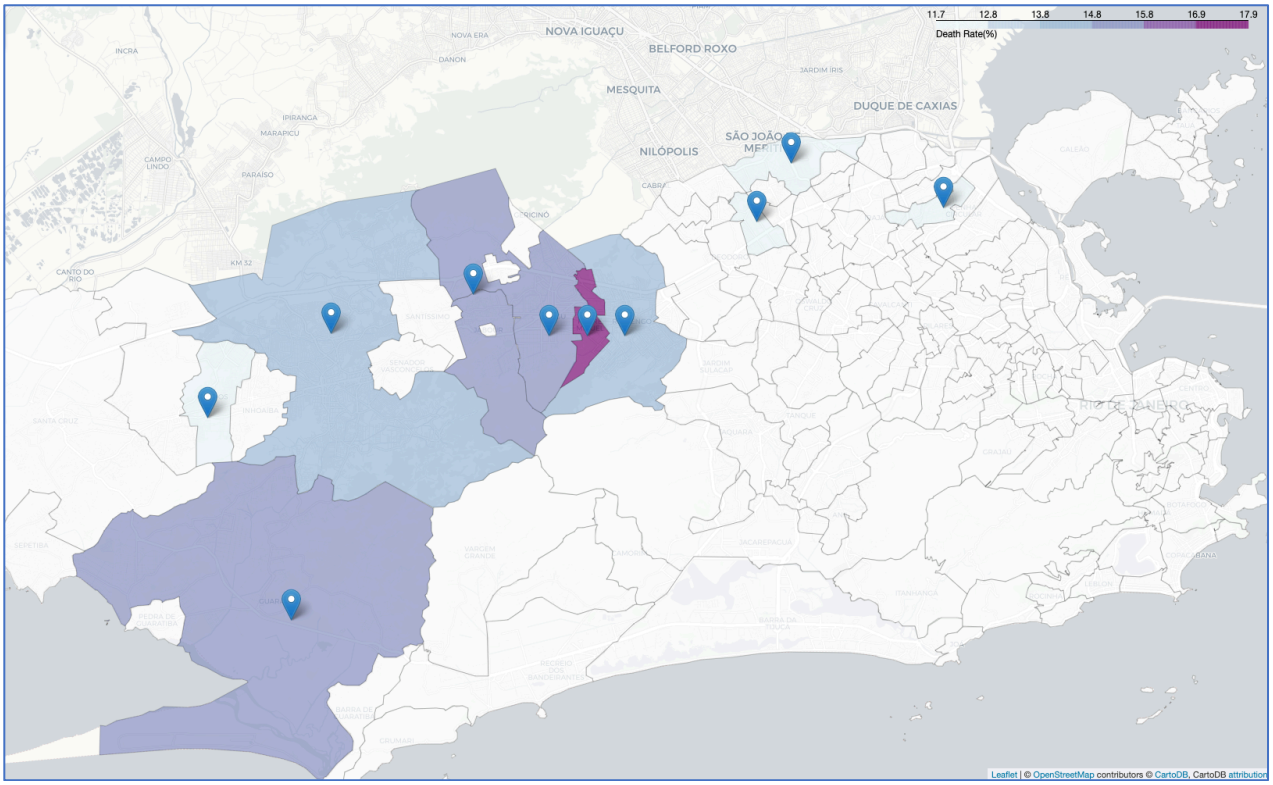

Figure 2 - Top ten regions with more than 1000 cases of COVID-19 shown in the Rio de Janeiro map

Observing the map, it is possible to infer a distinct characteristic from the COVID-19 death rate: the most affected neighborhoods are in the west zone of the city (the left part of the map).

Despite not having the proper data in this study, this region is well known as a low income, high density area. For diseases that spread rapidly, like COVID-19, those areas need special attention from the authorities.

Once the medical venues are gathered using the Foursquare API, we can fit the model with the K-Means algorithm.

3.2.Studying the clusters properties

Grouping the data by the cluster labels, we can investigate the death rate in each one:

	Cluster Labels	Death	Recovered	Total Cases	Death Rate	Recovery Rate	Death Rate(%)	Doctor's Office	Emergency Room	Hospital	Medical Center	Urgent Care Center	Total Infra
2	2	101.406250	1177.203125	1278.609375	0.088240	0.911760	8.824015	0.015625	0.375000	1.265625	0.000000	0.109375	1.765625
0	0	132.636364	1288.727273	1421.363636	0.078337	0.921663	7.833678	0.181818	0.727273	6.272727	0.000000	0.090909	7.272727
1	1	81.000000	1541.166667	1622.166667	0.052210	0.947790	5.221032	0.000000	1.500000	15.333333	0.333333	0.166667	17.333333

Figure 3 - Data grouped by the cluster label

It is easier to understand the characteristics of each cluster with a graph. First, we are going to visualize the average death rate:

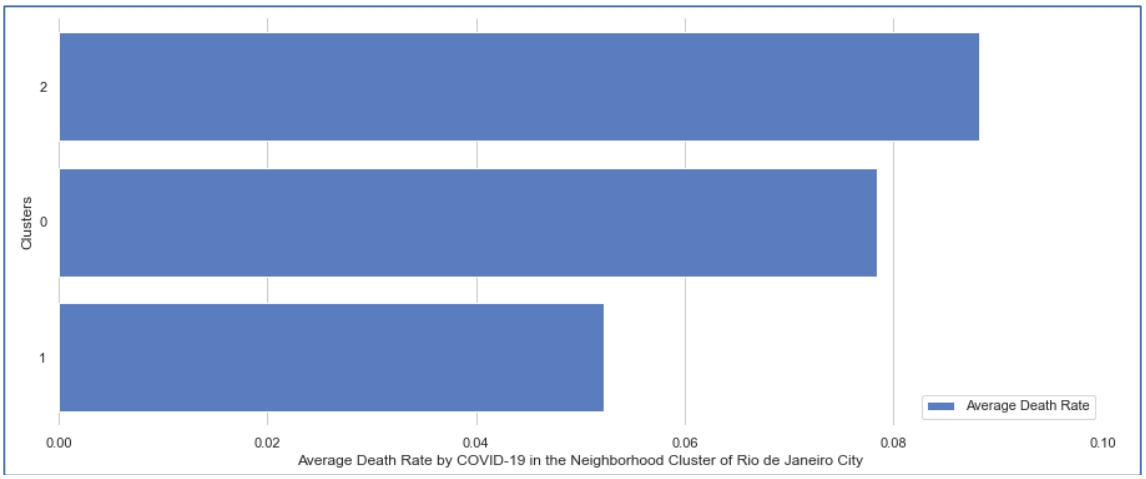

Figure 4 - Average Death Rate in each cluster

It is clear from the figure that cluster 2 has the larger average death rate (more than 8% of the total cases) while cluster 1 contains the most recovered cases (around 95% of the cases). Cluster 0 is in the middle with the death rate almost reaching the 8% of the total cases.

We can plot a similar graph considering the total medical infrastructure present in the cluster.

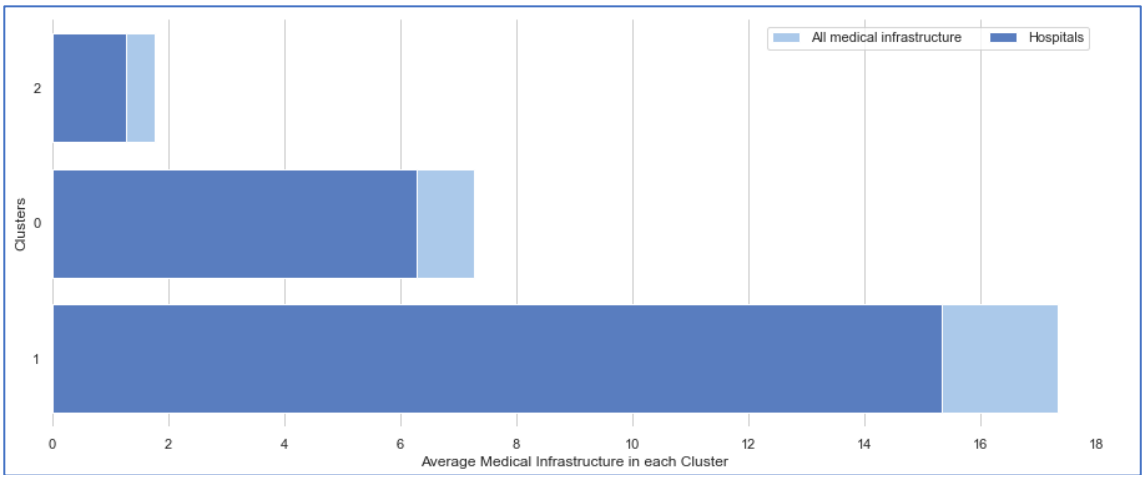

Figure 5 - Total medical infrastructure in the region and the quantity of hospitals

As expected, the clusters with better recovery rate have a larger medical infrastructure.

4. Results and Discussion

From the initial exploratory data analysis, we can see that:

- Cluster 2 has the larger death rate and worst medical infrastructure overall;
- Cluster 0 has a medium death rate and average medical infrastructure, with few hospitals;
- Cluster 1 has the smallest death rate and best medical infrastructure of all above.

The west zone of the city has the largest death rate by COVID-19. It consists of several neighborhoods, such as Campo Grande, Bangu, Realengo, Padre Miguel and many others.

It's important to remember that we are working only with the number of medical venues and not considering qualitative aspects and if they are prepared to deal with the pandemics.

Before reaching the conclusions, we are going to plot another Choropleth map with the cluster information obtained. Each marker in the map will represent a cluster. Their colors are the following:

- Cluster 0: being the cluster with the average medical infra, its color is orange;
- Cluster 1: since it has the best medical infra, its color is green;
- Cluster 2: the worst cluster overall, its color is red.

There are two neighborhoods with missing information regarding COVID-19 cases: Jabour and Vila Kennedy. They appear in the color black.

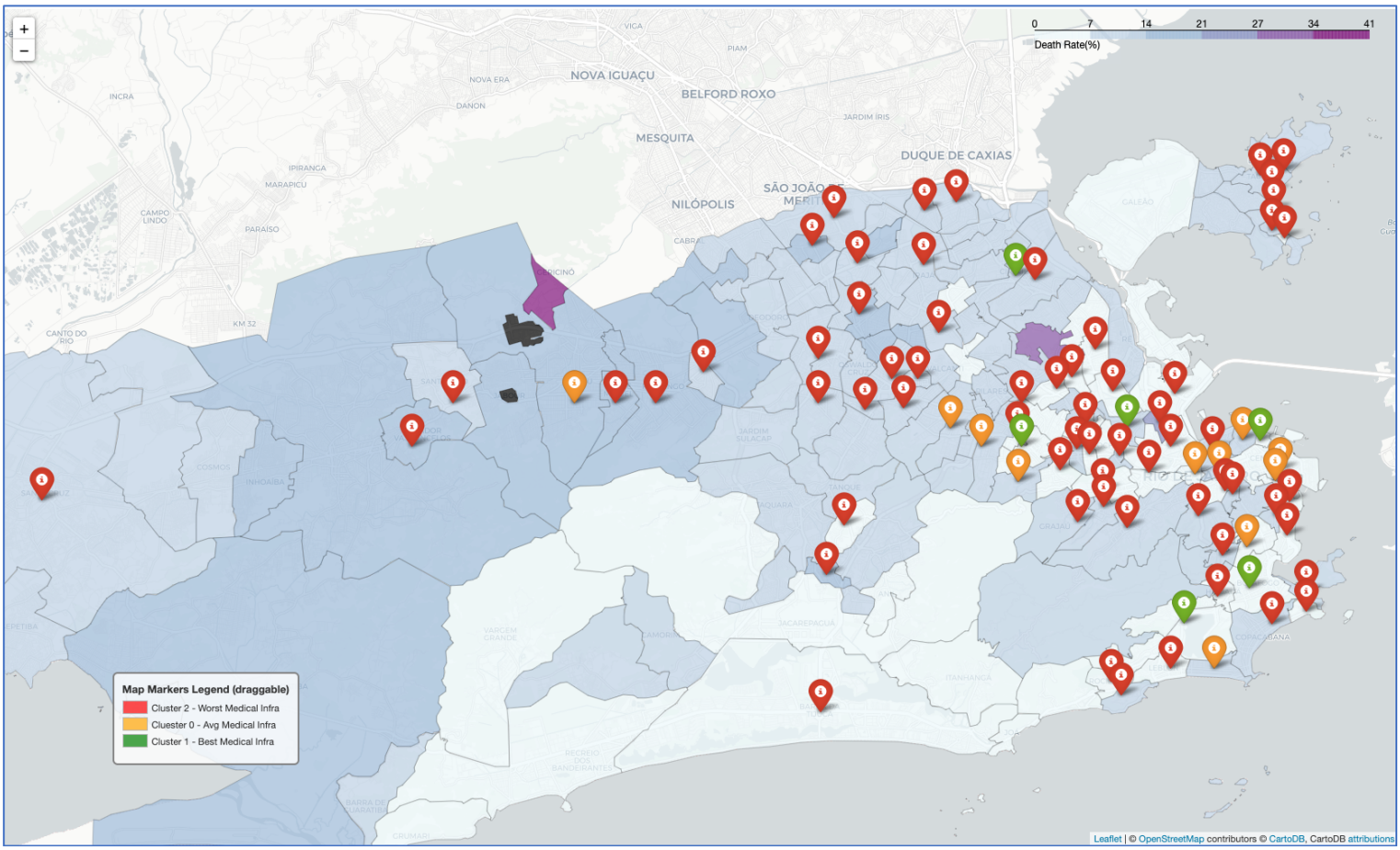

Figure 6 - Final Choropleth map with the cluster's identification

5. Conclusion

From this study, we can relate the importance of the medical infrastructure in the treatment and recovery of COVID-19 patients. (although some regions with good hospital infra have high death rates, others with precarious medical situation, have low death rates).

We know that there are many aspects that can result in a recovery or death of patients beyond the medical scope.

We should also consider that this analysis is quantitative and not qualitative. In fact, the number of hospitals itself is meaningless without knowing how it performs under the pandemic circumstances, its size, the number of employees available to work with cases of COVID-19, among other factors.

With this in mind, observing the map above, we can imply some conclusions:

- many of the light-colored areas (low death rate) with bad medical infrastructure have low demographic density, such as Vargem Grande and

Jacarepagua (areas with vast chunks of native rainforest), Galeao (the intl airport of Rio) and Cidade Universitaria (the region where the Federal University of Rio is situated);

- the largest death rates in the city are in the Gercino and Complexo do Alemao, where the first has a very low cases of COVID-19 registered (probably an outlier) and the latter is a famous slum, controlled by crime factions;
- worldwide known neighborhoods Ipanema and Copacabana, with reasonable medical infra in the area or nearby, have a high number of casualties probably because of the high average age of their residents, part of a risk group of the disease. This data is not shown in the study, but I think it is important to mention.
- All the west area of the city, places such as Padre Miguel, Senador Camara, Senador Vasconcelos, Campo Grande (the most populous neighborhood) and some others, support the results of the study since these are regions with high death rate and lacks medical infrastructure of all kinds.

So, in conclusion, if someone is looking for a place to live in Rio de Janeiro and the selection criteria is a safe place from the COVID-19 pandemic with good medical infrastructure, the entire west zone should be avoided. The best neighborhoods (part of cluster 1, in areas with low death rate) should be: **Jardim Botânico, Botafogo, Saude and Meier**.