

Report : Using Data Science to Analyze the Neighborhoods of Rio de Janeiro

Wanderson Torres

April, 2021

Analyzing the Neighborhoods of Rio de Janeiro, RJ

Author: Wanderson Torres

1. Introduction:

1.1 Background

The coronavirus COVID-19 is an infectious disease that has quickly spread to over 200 countries in a matter of weeks, disrupting our daily routines and our lives as we knew them. According to Worldometers website, there has been over 147 million total cases and 3 million total deaths in the world, as of April 25, 2021. In Brazil alone, there has been over 14 million total cases and close to 400,000 deaths. Because the virus is thought to spread mainly from person to person through respiratory droplets, many countries, including Brazil, have asked their citizens to quarantine themselves since March 2020 and to limit travel and physical human contact outside of their household.

The emergence of the coronavirus disease 2019 (COVID19), has devastated economies and caused unprecedented challenges to healthcare and food systems around the world. Globally, a lot of jobs have gone remote, thus giving people an option to move farther away from the office to a safety area. We have seen a trend where people are moving away from crowded and expensive neighborhoods to the suburbs or even to other cities/states where there is more secure and more protected against the virus.

1.2 The problem

I live in the city of Rio de Janeiro and due to the increase in the number of deaths and the occupation of the hospitals beds reaching their maximum occupancy, I would like to conduct a study where it points out areas of lower COVID risk and high options for health treatment in general.

The aim of the project is to apply the skills learned in the Coursera course to find the safest neighborhood in Rio de Janeiro,

surrounded by hospitals, drugstores, clinics and so on. This will be determined by analyzing the number of cases of covid, deaths and the profile of the neighborhood's population, clustering neighborhoods using k-means and exploring on the map the top common healthy venues in the safest neighborhoods.

1.3 Interest

This exercise may also be of interest to anyone who is facing this pandemic and is concerned about the health issue.

It can be interesting for those who want to move to the city of Rio de Janeiro. By segmenting and clustering neighborhoods in Rio, and analyzing the Health Services we can determine the most suitable neighborhood we most want to live in.

2. The Data

In this study, I worked with 5 main data sources:

2.1 Neighborhoods Dataset

I also obtained a list of Rio de Janeiro's districts and neighborhoods by web scraping a Wikipedia webpage using BeautifulSoup and after some data cleaning steps, we have the dataset bellow.

	City	Zone	District	Neighborhood
0	Rio de Janeiro	Central	Centro Histórico e Zona Portuária	São Cristóvão
1	Rio de Janeiro	Central	Centro Histórico e Zona Portuária	Benfica
2	Rio de Janeiro	Central	Centro Histórico e Zona Portuária	Caju
3	Rio de Janeiro	Central	Centro Histórico e Zona Portuária	Catumbi
4	Rio de Janeiro	Central	Centro Histórico e Zona Portuária	Centro

However, this dataset lacked the geographical coordinates. So, I used **geocoder python library** to obtain the latitude and longitude coordinates for each Rio de Janeiro neighborhood.

Following below plot with **164** neighborhoods' location overview.

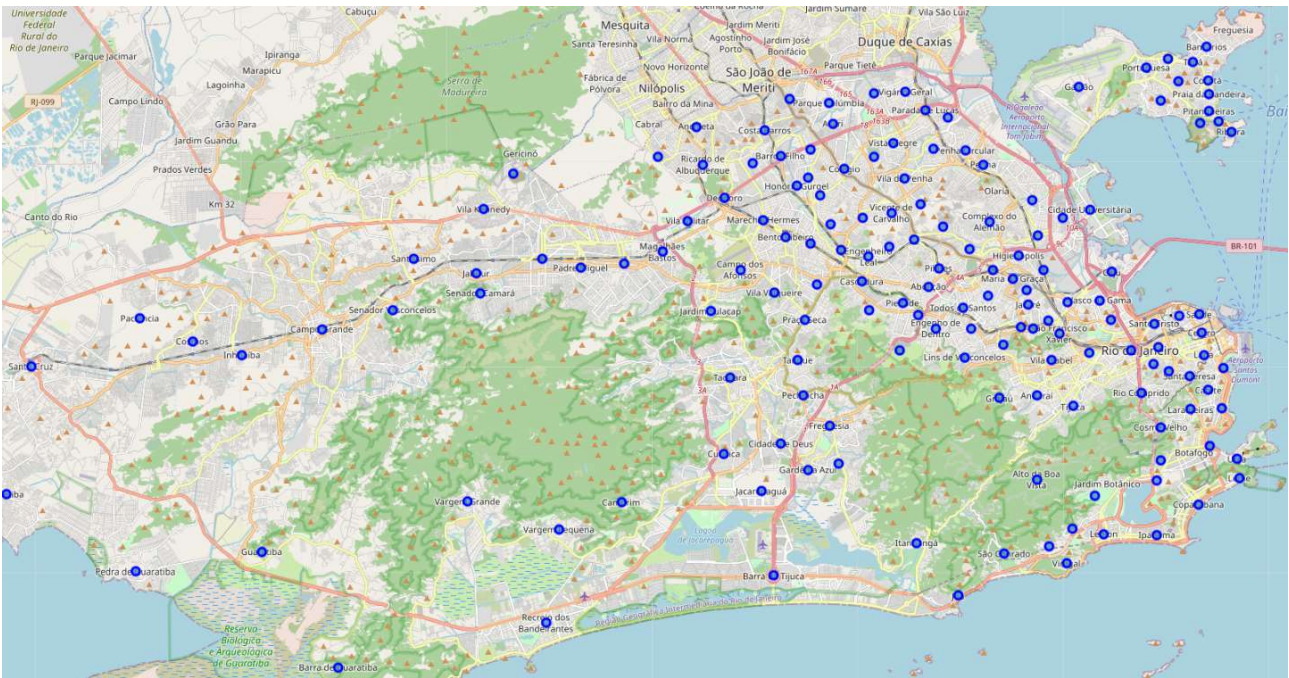


Fig. 1: City of Rio de Janeiro and all neighborhoods

2.2 COVID Data

The Rio Covid Panel is an initiative of the city of Rio de Janeiro. It's an open source and available for download directly from the Data Rio Website(<https://www.data.rio/app/painel-rio-covid-19>). It contained the features about COVID cases: neighborhood of residence, case evolution, age group and sex.

	bairro_resid_estadia	sexo	faixa_etaria	evolucao	raca_cor
0	ABOLICAO	F	De 70 a 79	OBITO	Preta
1	ABOLICAO	F	De 90 a 99	OBITO	Parda
2	ABOLICAO	F	De 70 a 79	OBITO	Branca
3	ABOLICAO	M	De 40 a 49	OBITO	Preta
4	ABOLICAO	M	De 70 a 79	OBITO	Parda

2.2.1 Data Cleaning in Covid Dataset

The covid dataset has 243,790 cases in 164 neighborhoods in Rio de Janeiro. After some investigations, I discover that some records were with the names of the neighborhoods as being undefined or outside the municipality.

The total number of registrations of this type was almost 20,000 cases. I decided to delete all cases for 2 reasons: First, I

consider, for this academic work, there is enough data for clustering and second, the data taken represents less than 10% of the total.

2.2.1 Feature Selection

After data cleaning, there were 219.405 samples and 4 categorical features in the data which could be transformed in more than 15 new features. Upon examining the meaning of each feature, it was clear that there was some redundancy in the features.

The same covid case is represented in all columns features. For example, there was a feature of sex of the person, and another feature of the age group and race of the same person. These three features contained very similar information (they refer to the same person), with the difference being only the characteristics.

These features are problematic for two reasons: (1) The number of total cases will be duplicated in three features. (2) To analyze this type of information and its relationships with covid cases, specialists in sociology and medicine are needed, which would make the scope of this work unfeasible for the required term.

In order to fix this, I decided to drop all features related to the profile of the patients and kept the evolution of the case of covid.

After discarding redundant features, I inspected the correlation of independent variables, and 1 feature was selected. This feature will be transformed in 3 new features and will be merged with the population.

The dataset result contains 219.405 samples and 1 categorical feature.

Neighborhood evolution		
0	ABOLICAO	OBITO
1	ABOLICAO	OBITO
2	ABOLICAO	OBITO
3	ABOLICAO	OBITO
4	ABOLICAO	OBITO

2.3 Population of Rio de Janeiro

I have got the number of habitants in each neighborhood in Rio de Janeiro from the Data Rio Website (<https://www.data.rio/search?groupIds=0f4009068ec74e17b25eb3e70891b95f&sort=-modified>). It's an open-source information and available to download for free.

After some data cleaning steps, we have all the numbers of habitants of all the 164 neighborhoods of the city of Rio de Janeiro.

	Neighborhood	Population
0	COPACABANA	161031
1	LEBLON	50648
2	IPANEMA	47017
3	FLAMENGO	55047
4	MEIER	54811

2.4 Foursquare Venue Data

First of all, we are interested in Health Services Points. On the Foursquare Developer website, we can check and choose the categories that are part of the health services. Below are the ones I chose for this study:

Category_Name	Category_ID
Medical Center	4bf58dd8d48988d104941735
Acupuncturist	52e81612bcbc57f1066b7a3b
Alternative Healer	52e81612bcbc57f1066b7a3c
Chiropractor	52e81612bcbc57f1066b7a3a
Dentist's Office	5744ccdf4b0c0459246b4d6
Doctor's Office	4bf58dd8d48988d177941735
Emergency Room	4bf58dd8d48988d194941735
Eye Doctor	522e32fae4b09b556e370f19
Hospital	4bf58dd8d48988d196941735
Hospital Ward	58daa1558bbb0b01f18ec1f7
Maternity Clinic	56aa371be4b08b9a8d5734ff
Medical Lab	4f4531b14b9074f6e4fb0103
Mental Health Office	52e81612bcbc57f1066b7a39
Nutritionist	58daa1558bbb0b01f18ec1d0
Physical Therapist	5744ccdf4b0c0459246b4af
Rehab Center	56aa371be4b08b9a8d57351d

Urgent Care Center	56aa371be4b08b9a8d573526
Drugstore	5745c2e4498e11e7bccabdbd
Health & Beauty Service	54541900498ea6ccd0202697
Massage Studio	52f2ab2ebcbc57f1066b8b3c
Medical Supply Store	58daa1558bbb0b01f18ec206
Pharmacy	4bf58dd8d48988d10f951735
Spa	4bf58dd8d48988d1ed941735
Supplement Shop	5744ccdf4b0c0459246b4cd
Health & Beauty Service	54541900498ea6ccd0202697
Health Food Store	50aa9e744b90af0d42d5de0e

So, with the categories, we will use RESTful API calls to retrieve data about the venues in different areas. As mentioned, we will use Foursquare API to explore all neighborhoods in Rio de Janeiro city.

To get venue information in each neighborhood, I called the [Foursquare API] (<https://foursquare.com/developers/apps>). Following below an example of a response from the API.

```
{'categories': [{ 'id': '4d4b7104d754a06370d81259',
'name': 'Arts & Entertainment',
'pluralName': 'Arts & Entertainment',
'shortName': 'Arts & Entertainment',
'icon': { 'prefix': 'https://ss3.4sqi.net/img/categories_v2/arts_entertainment/default_',
'suffix': '.png'},
'categories': [{ 'id': '56aa371be4b08b9a8d5734db',
'name': 'Amphitheater',
'pluralName': 'Amphitheaters',
'shortName': 'Amphitheater',
'icon': { 'prefix': 'https://ss3.4sqi.net/img/categories_v2/arts_entertainment/default_',
'suffix': '.png'},
'categories': []}],
{'id': '4fceeal71983d5d06c3e9823',
'name': 'Aquarium',
'pluralName': 'Aquariums',
'shortName': 'Aquarium',
'icon': { 'prefix': 'https://ss3.4sqi.net/img/categories_v2/arts_entertainment/aquarium_',
'suffix': '.png'},
'categories': []}]}
```

Because of the API restrictions I had to split the query into 5 parts. In each query, I'd like to know the number of venues in each category in a 2000m radius around each neighborhood of Rio. This gave me 5 datasets containing the venue name, latitude and longitude coordinates of the venue location, and the venue category.

After I merge all of them. The result is 2,694 rows and 5 columns.









	Neighborhood	Latitude	Longitude	category_name	venues
0	São Cristóvão	-22.899318	-43.221935	Medical Center	30
1	São Cristóvão	-22.899318	-43.221935	Alternative Healer	2
2	São Cristóvão	-22.899318	-43.221935	Dentist's Office	2
3	São Cristóvão	-22.899318	-43.221935	Doctor's Office	9
4	São Cristóvão	-22.899318	-43.221935	Emergency Room	5

2.4 Geodata of Rio de Janeiro

I downloaded georeferenced information about a map of the neighborhoods in Rio de Janeiro. The API can be found on the DATA RIO website (https://opendata.arcgis.com/datasets/dc94b29fc3594a5bb4d297bee0c9a3f2_15.geojson). It contains the necessary data for making and painting the choropleth map using Folium.

Showing 1 to 10 of 163

Hint: Filter columns using 

 OBJECTID	 BaseGeo_DBO_LimiteBairro_AREA	 NOME	 REGIAO_ADM	 AREA_PLANE	 CODBAIRRO	 CODRA	 CODBNUM
325	1705684.50390625	Paquetá	PAQUETA	1	013	21	Paquetá
326	4056402.76611328	Freguesia (Ilha)	ILHA DO GOVERNADOR	3	098	20	Freguesia
327	978046.54248047	Bancários	ILHA DO GOVERNADOR	3	097	20	Bancários
328	18957422.15185547	Galeão	ILHA DO GOVERNADOR	3	104	20	Galeão
329	1672545.75097656	Tauá	ILHA DO GOVERNADOR	3	101	20	Tauá

3. Exploratory Data Analysis

It's time to look deeper into the data. It has been created a descriptive analysis and a box plot in order to get an overview of all neighborhoods.

Most of Covid Cases in Rio are "Recuperado" (Recovered).

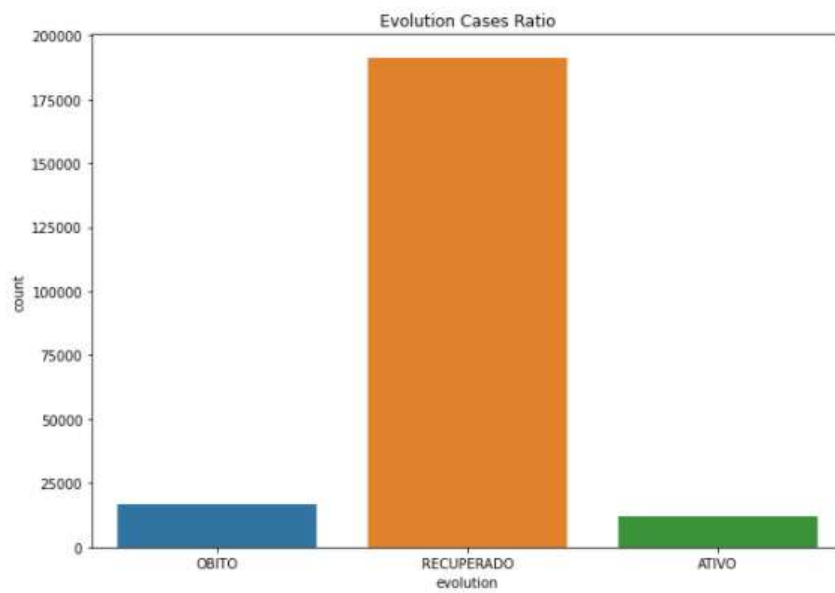
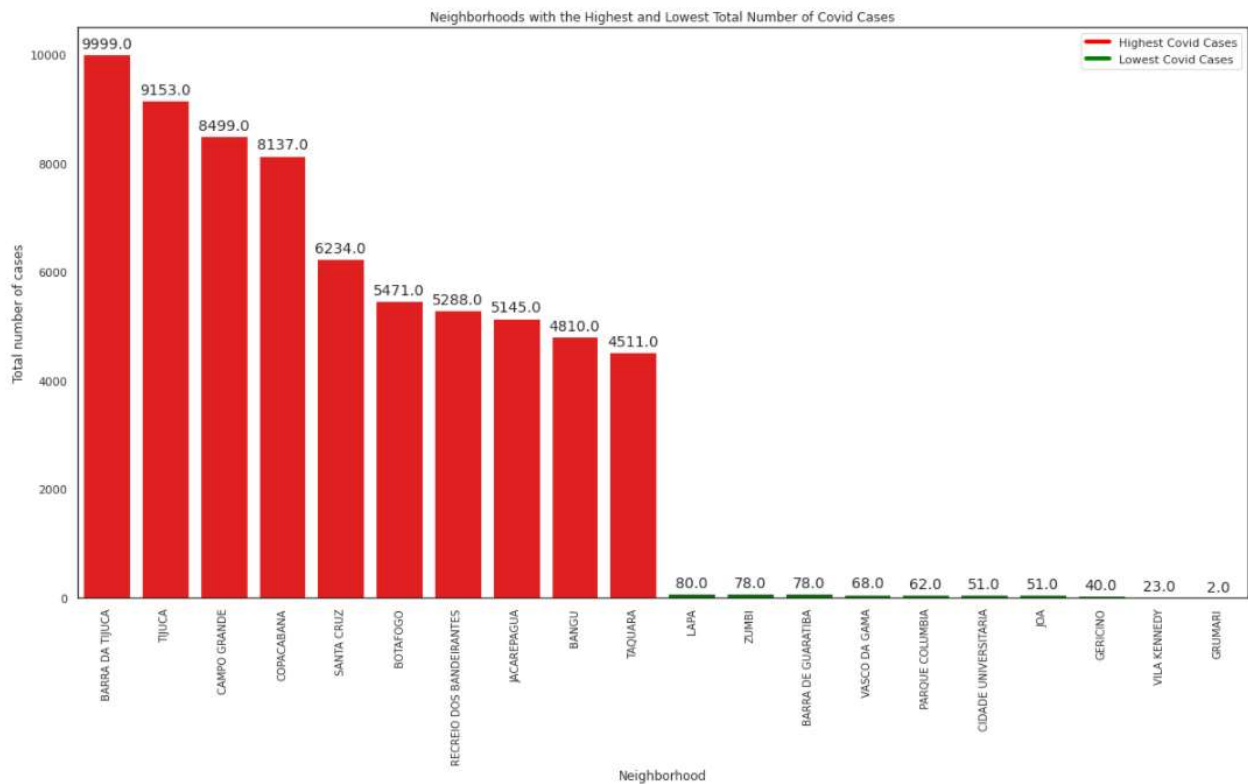


Fig 2. Number of Covid Cases by Evolution in Rio de Janeiro

After I grouped the sum of evolution cases by neighborhood.

evolution	ATIVO	OBITO	RECUPERADO	Total
Neighborhood				
BARRA DA TIJUCA	674.0	422.0	8903.0	9999.0
TIJUCA	400.0	615.0	8138.0	9153.0
CAMPO GRANDE	424.0	1037.0	7038.0	8499.0
COPACABANA	690.0	577.0	6870.0	8137.0
SANTA CRUZ	149.0	461.0	5624.0	6234.0
BOTAFOGO	284.0	203.0	4984.0	5471.0
RECREIO DOS BANDEIRANTES	351.0	255.0	4682.0	5288.0
JACAREPAGUA	302.0	265.0	4578.0	5145.0
BANGU	199.0	693.0	3918.0	4810.0
TAQUARA	334.0	357.0	3820.0	4511.0
REALENGO	153.0	552.0	3235.0	3940.0
CENTRO	111.0	139.0	3449.0	3699.0
VILA ISABEL	141.0	228.0	3164.0	3533.0
LEBLON	246.0	146.0	2967.0	3359.0

I also found the 10 districts with the highest total number of COVID cases and the lowest total number of cases. I then concatenated these 2 lists and created a bar chart to better visualize the disparity.



3.1 One Hot Encoding and Merge Datasets

To continue the analysis, we must add information of the population by neighborhood.

Before, we're going to first use one-hot encoding to turn our categorical variable (the evolution case) in our dataset into a binary vector.

	Neighborhood	ATIVO	OBITO	RECUPERADO
0	ABOLICAO	0	1	0
1	ABOLICAO	0	1	0
2	ABOLICAO	0	1	0
3	ABOLICAO	0	1	0
4	ABOLICAO	0	1	0

And now we merge with the Population dataset.

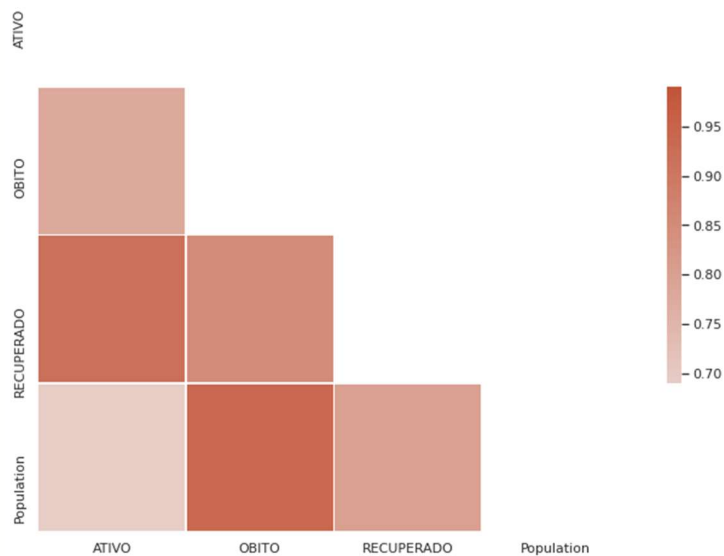
	Neighborhood	ATIVO	OBITO	RECUPERADO	Population
0	ABOLICAO	38.0	47.0	469.0	12492.0
1	ACARI	11.0	53.0	239.0	30082.0
2	AGUA SANTA	15.0	21.0	218.0	9632.0
3	ALTO DA BOA VISTA	21.0	33.0	311.0	10277.0
4	ANCHIETA	60.0	104.0	1097.0	61217.0

3.2 Analyzing the Correlation

First, I create the correlation Matrix

	ATIVO	OBITO	RECUPERADO	Population
ATIVO	1.000000	0.777573	0.922917	0.689583
OBITO	0.777573	1.000000	0.851842	0.939468
RECUPERADO	0.922917	0.851842	1.000000	0.799310
Population	0.689583	0.939468	0.799310	1.000000

And after the Heatmap, is a graphical representation of the correlation matrix. The higher the coefficient, the greater the correlation.



So, we can conclude with the correlation matrix above that the variable Population (number of inhabitants) has a strong relationship with the variable Active (Active Covid Cases) - 0.69

and has a very strong relationship with the variable DEATH (Deaths) - 0.94 and with the variable RECOVERED (Recovered from COVID) - 0.80.

4. Feature Scaling

After select and clean all the features and data we have, it's time for feature scaling. It is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

The specific method that we use for this problem is Feature standardization. It makes the values of each feature in the data have zero-mean (when subtracting the mean in the numerator) and unit-variance. This method is widely used for normalization in many machine learning algorithms.

```
array([[ -3.57460187e-01, -4.12304973e-01, -4.85315471e-01,
        -6.20525467e-01],
       [ -6.28812297e-01, -3.67492222e-01, -6.42623342e-01,
        -2.67985566e-01],
       [ -5.88611984e-01, -6.06493562e-01, -6.56986235e-01,
        -6.77845769e-01],
       [ -5.28311515e-01, -5.16868059e-01, -5.93379139e-01,
        -6.64918638e-01],
       [ -1.36358468e-01,  1.34161632e-02, -5.57965870e-02,
        3.56024089e-01],
       [ -3.58576862e-02, -1.65834842e-01,  7.48373411e-02,
        -3.02948320e-03],
       [  4.26445908e-01, -6.87405474e-02,  2.12994689e-01,
        -3.37992494e-01],
       [ -4.68011046e-01, -5.39274435e-01, -5.66021248e-01,
        -5.95052004e-01],
       [  1.26060239e+00,  4.41253457e+00,  1.87361865e+00,
        4.48911851e+00],
```

5. Cluster the Neighborhoods

5.1 Find the best number of Clusters

5.1.1. Elbow Method

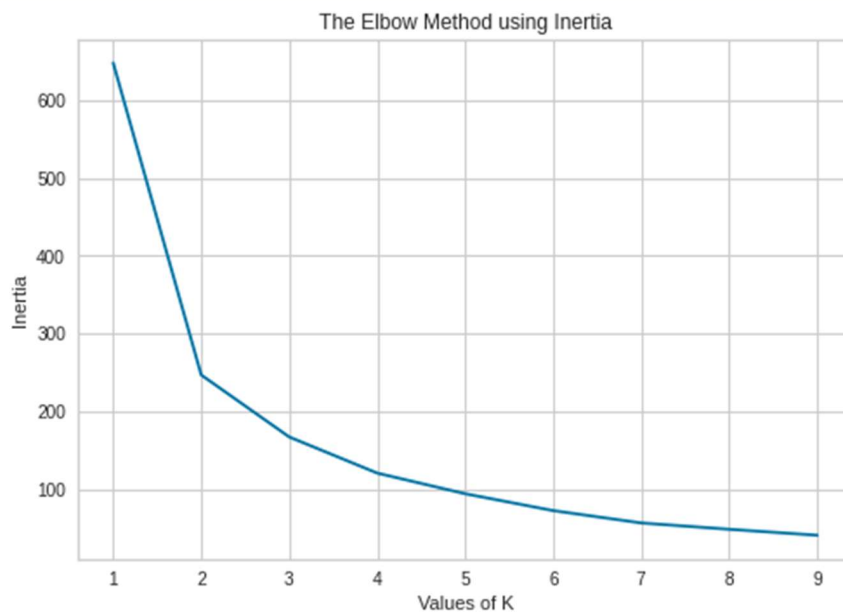
In clustering, use the "elbow" as a cutoff point is a common heuristic in mathematical optimization to choose a point where means choose a number of clusters so that adding another cluster doesn't give much better modeling of the data.

The intuition is that increasing the number of clusters will naturally improve the fit (explain more of the variation), since

there are more parameters (more clusters) to use, but that at some point this is over-fitting, and the elbow reflects this.

In practice there may not be a sharp elbow, and as a heuristic method, such an "elbow" cannot always be unambiguously identified.

In our case, after 5 or 6 clusters, more number of clusters doesn't give much better modeling of the data.



5.1.2 Silhouette Score

Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other.

The value of the Silhouette score varies from -1 to 1. If the score is 1, the cluster is dense and well-separated than other clusters. A value near 0 represents overlapping clusters with samples very close to the decision boundary of the neighboring clusters. A negative score [-1, 0] indicates that the samples might have got assigned to the wrong clusters.

In our case, considered the best clusters [2-6] because the scores are near or above 0.5:

```
N_cluster: 2, score: 0.7882471425894928  
N_cluster: 3, score: 0.5106348057892676  
N_cluster: 4, score: 0.515152346058788  
N_cluster: 5, score: 0.4977375527013884
```

N_cluster: 6, score: 0.5009512483000026
N_cluster: 7, score: 0.46409363723427893

5.1.3 Silhouette Visualizer

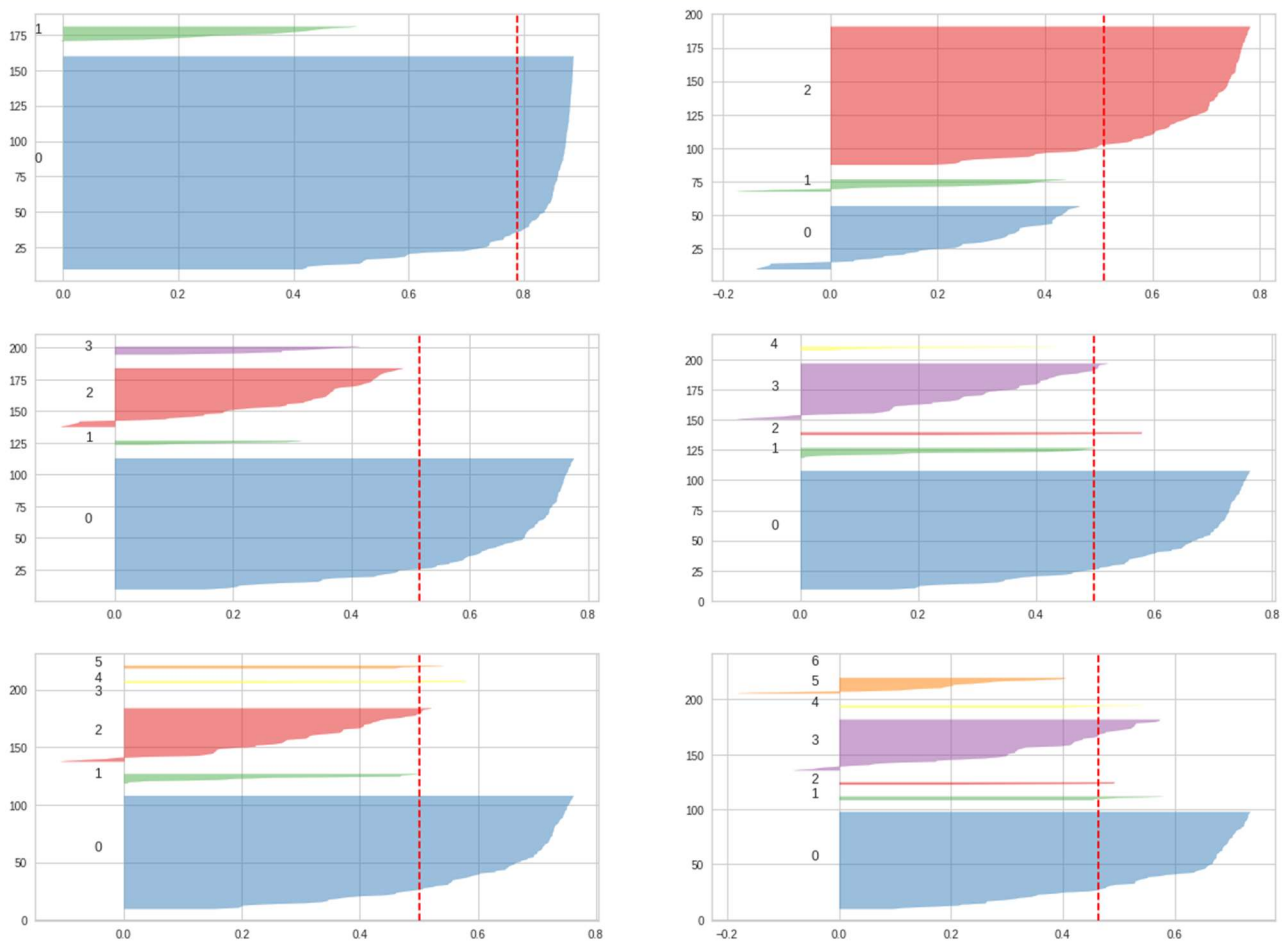
Here is the Silhouette Visualizer analysis done on the above plots to select an optimal value for n_clusters.

The value of n_clusters as 2, 3, 4, 5, 7 looks to be suboptimal for the given data due to the following reasons:

- * Presence of clusters with below-average silhouette scores
- * Wide fluctuations in the size of the silhouette plots.

The value of 6 for n_clusters look to be the optimal one. The silhouette score for each cluster is above average silhouette scores, despite the fluctuation in size is not similar. The thickness of the silhouette plot representing each cluster also is a deciding point. The value of 7 clusters has more negative samples that the 6 clusters.

Thus, one can select the optimal number of clusters as 6.



5.2 Clustering

Use 6 clusters as optimal number of clusters for the K-Means clustering.

The K-means is an unsupervised learning algorithm that will be used for this study to cluster the neighborhoods based on the features.

All the modifications in our dataset have already done. A colored-coded map of clusters was then created for better visualization.

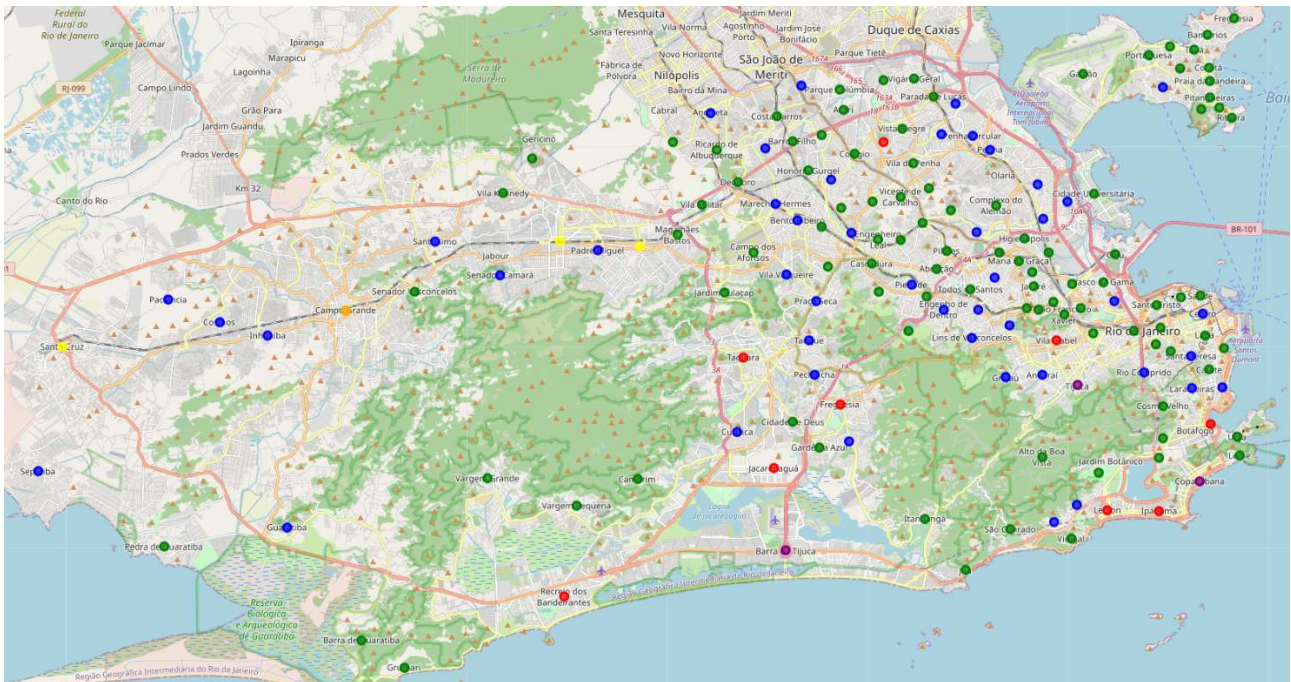


Fig.3 - Map of Clusters

6. Analyzing the Clustering

Now that the AI has grouped Rio de Janeiro's Neighborhoods into clusters, I want to look deeper into each cluster. Maybe even figure out why the AI grouped the clusters the way it did and find the best cluster for me to move.

Previous analysis, I can conclude:

- **Cluster 0 (green)**
 - Many neighborhoods
 - Neighborhood's population is low
 - Low number of COVID cases
- **Cluster 1 (red)**
 - Few Neighborhoods

- Neighborhood's population is high
- High number of COVID cases
- **Cluster 2 (blue)**
 - Many Neighborhoods
 - Middle Neighborhood's population
 - Middle number of COVID cases
- **Cluster 3 (orange)**
 - 1 Neighborhood
 - Highest Neighborhood's population
 - Middle number of COVID cases
 - High number of Deaths
- **Cluster 4 (purple)**
 - Few Neighborhoods
 - High Neighborhood's population
 - High number of COVID cases
 - Low number of Deaths
- **Cluster 5 (yellow)**
 - Few Neighborhoods
 - High Neighborhood's population
 - Middle number of COVID cases
 - High number of Deaths

6. Create the COVID RISK INDEX (CRI)

In my analysis, it is still difficult to understand and choose one cluster to explore and consider with low risk to COVID. So, I decided to create an index.

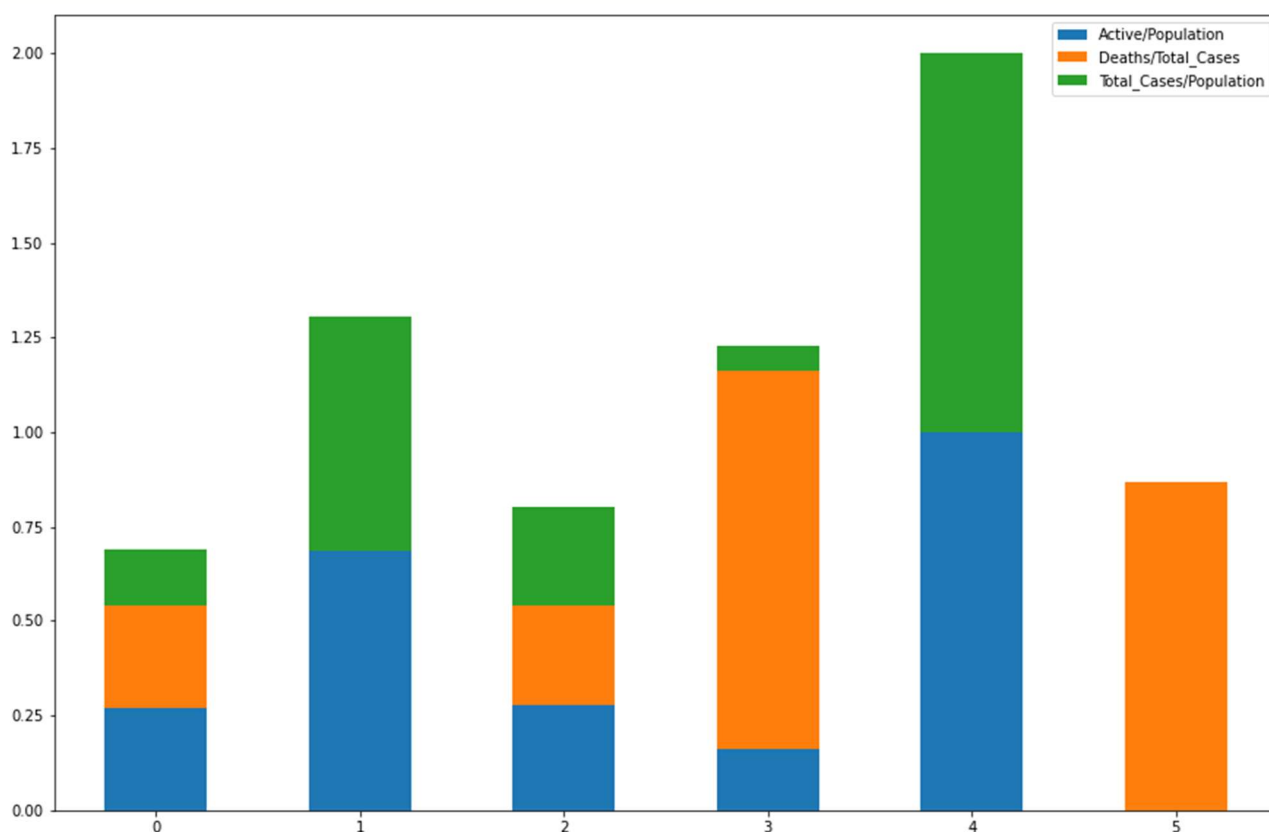
The covid risk index is an indicator that considers 3 variables:

1. The number of active COVID cases by the population.
2. Number of deaths by COVID by the total number of COVID cases.
3. Total number of COVID cases by the population.

$$\text{CRI} = 1 + 2 + 3$$

After normalize the variables and calculating the Covid Risk Index, the dataframe and plot are below:

	Cluster	ATIVO	OBITO	RECUPERADO	Population	Active/Population	Deaths/Total_Cases	Total_Cases/Population	IRC
	0	2888	3905	44441	1942240	0.269086	0.271581	0.148907	0.69
	1	2265	2121	31473	842908	0.685293	0.000058	0.619535	1.30
	2	4076	6174	71290	2697684	0.277402	0.263605	0.260924	0.80
	3	424	1037	7038	361207	0.160506	1.000000	0.065941	1.23
	4	1764	1614	23911	490733	1.000000	0.000000	1.000000	2.00
	5	501	1706	12777	704639	0.000000	0.870214	0.000000	0.87



Now, with CRI, I can better assess the risk of COVID per cluster.

If I'd like to consider the total number of the index, cluster 4 would be the worst cluster in relation to risk and cluster 0 would be the best.

But the analysis can be done separately between the variables.

For example, if I am concerned about the lethality of people who take covid, I should consider the orange bands on the chart above and the larger the band, the greater the risk. In this case, clusters 3 and 5 are the ones that offer the highest risk and clusters 1 and 4 are the ones that offer the lowest risk.

In my case, I will consider the total value of the index, thus I create a table with the risk value and cluster and classified:

Cluster	CRI
Cluster 0	0.69
Cluster 2	0.80
Cluster 5	0.87
Cluster 3	1.23
Cluster 1	1.30
Cluster 4	2.00

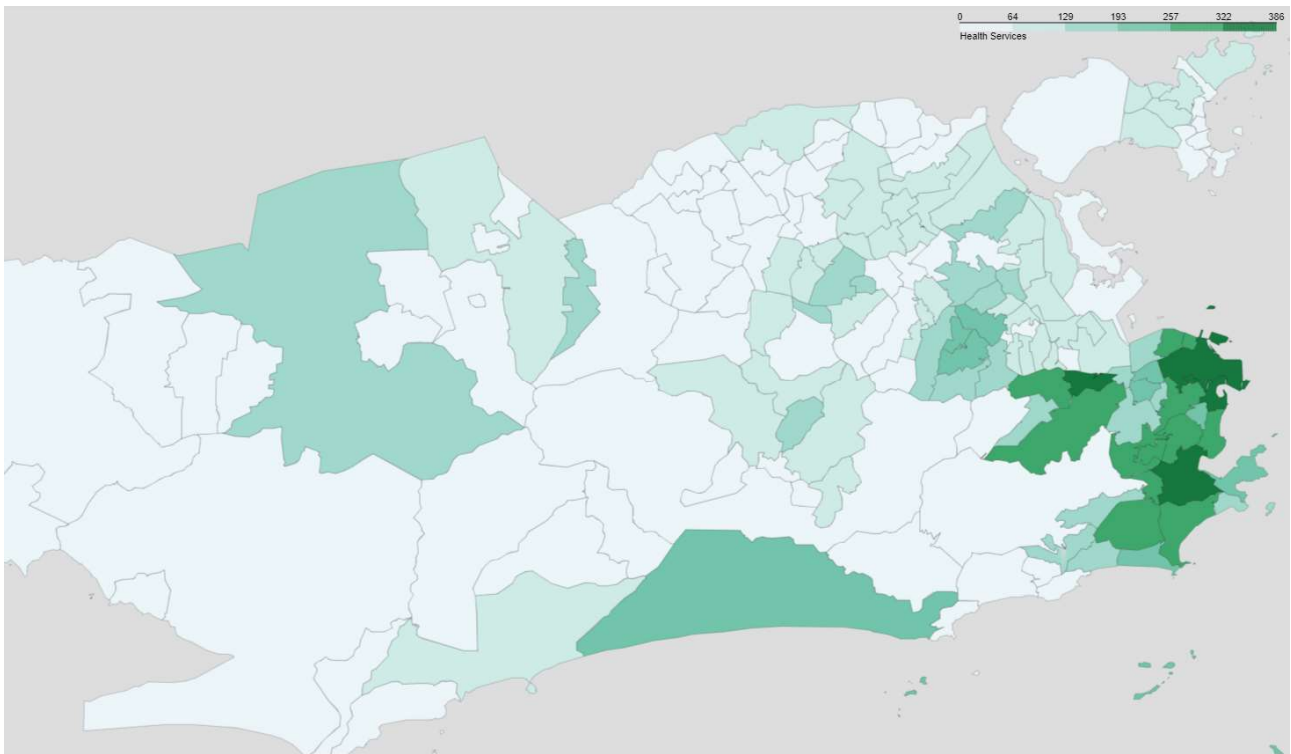
7. Explore the neighborhoods with health services in Rio de Janeiro using Foursquare API

The analysis is not over. Concerned about health, I want to choose a place with a low risk of COVID, but that offers a high amount of health services such as clinics, hospitals, pharmacies, doctors' offices, etc.

For this, I used the Foursquare API to get the total points of Health Services by Neighborhood.

Index	CODBAIRRO	Neighborhood	venues	Level_labels
0	0	020	Botafogo	386 High-1 Level HV
1	1	016	Glória	358 High-1 Level HV
2	2	035	Maracanã	355 High-1 Level HV
3	3	161	Lapa	353 High-1 Level HV
4	4	005	Centro	336 High-1 Level HV

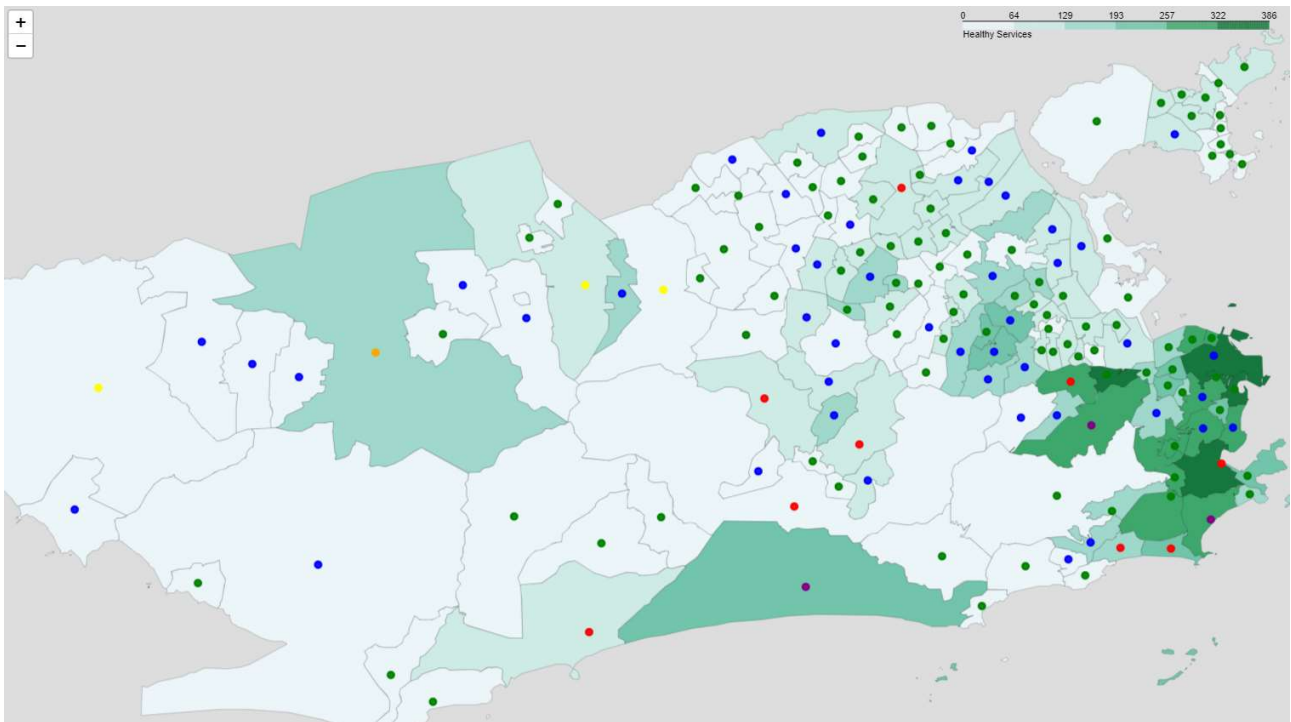
After that I merged the code of each neighborhood with the geojson information dataset to create the Choropleth Map using **FOLIUM python library**. The map is below.



As we can see the in the map, the city center and the south zone of Rio de Janeiro are the regions that offer the largest number of health service points.

7. Choose a safer place to live

If we combined COVID risk clustering and health service offerings, we could look at the map and choose the place to live considering these factors.



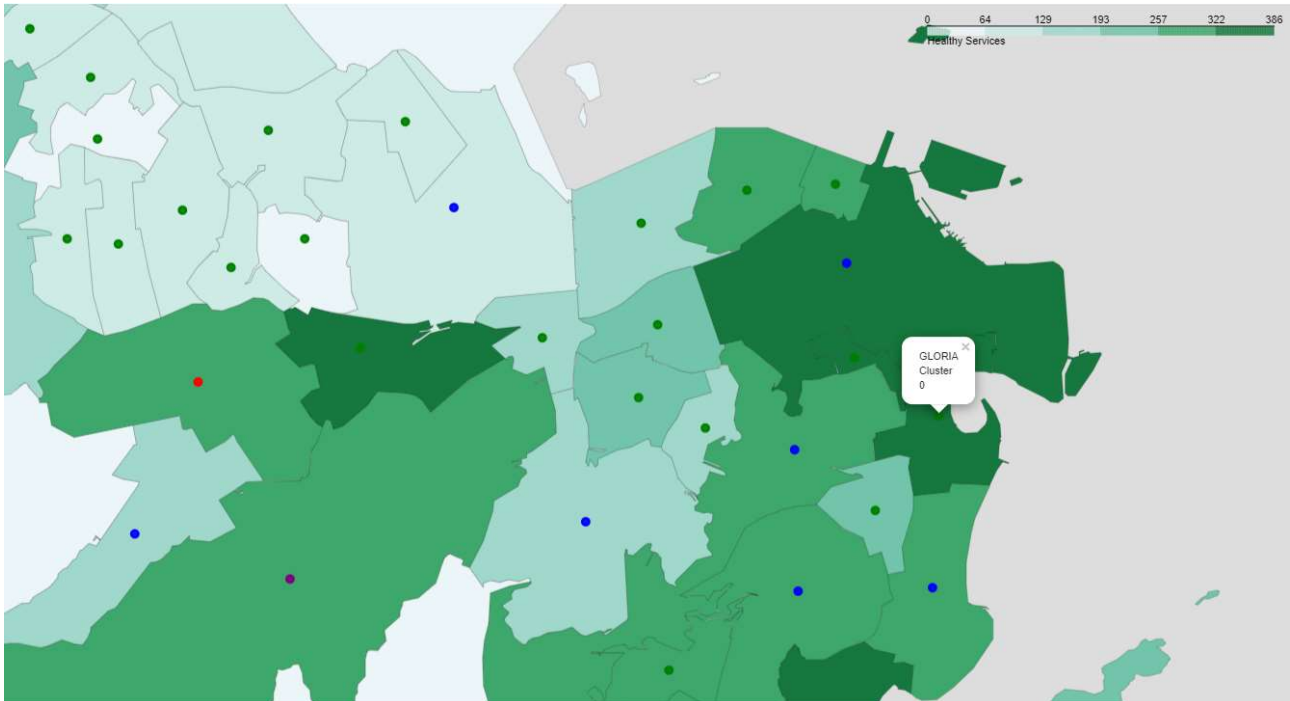
Remembering that the colors represent the 6 clusters, and the risk scale is in accordance with the CRI (Covid Risk Index) as we can see in the table below:

Cluster	Color	CRI
Cluster 0	green	0.69
Cluster 2	blue	0.80
Cluster 5	yellow	0.87
Cluster 3	orange	1.23
Cluster 1	red	1.30
Cluster 4	purple	2.00



So what we are looking for on the map are **dark green areas**, as they have a high supply of health services and **green or blue dots** because they offer less risk compared to COVID.

Getting deeper into dark green areas, I can see 2 green dots that represents low risk. One is Lapa Neighborhood and the other is Glória.



These 2 neighborhoods have a high number of health services like clinics, hospitals, pharmacies, medical's office etc... and have a low risk consider the CRI calculus.

8. Discussion

As I mentioned before, Rio de Janeiro is a big city with a high population density in a large area. The total number of measurements and population densities of the 164 neighborhoods in total can vary. As there is such a complexity, very different approaches can be tried in clustering and classification studies. Moreover, it is obvious that not every classification method can yield the same high quality results for this metropol.

We also have to consider covid underreporting regions mainly in poorer areas of the city and slums. This may have a bias in the analysis.

I used the Kmeans algorithm as part of this clustering study but I can use other algorithms and compare the results.

Foursquare in Brazil is little used and therefore the data may also be sent to regions where they are most used.

I ended the study by visualizing the data and clustering information on the Rio de Janeiro map. In future studies, could include more features like age group, sex and race and analyze the clusters with Health professionals.

9. Conclusion

Six clusters were identified by K-means algorithm. The Covid risk index (CRI) were created to compare the clusters. All kind of Health Services were identified, added and plotted on the map. It's showed that Rio de Janeiro has more venues concentration in downtown and South zone. Combining it with CRI and analyzing on the map, I have now 2 neighborhoods to look for apartment for rent: GLÓRIA and LAPA.