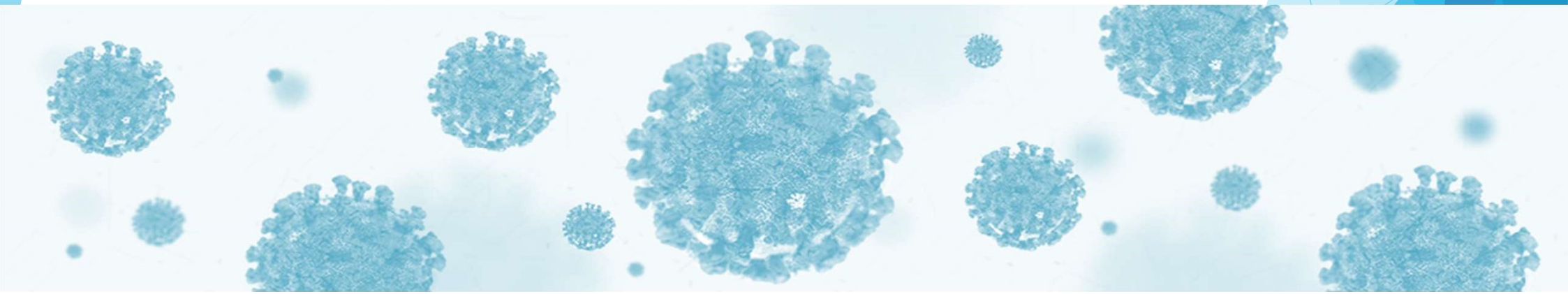
Using Data Science to Analyze the Neighborhoods of Rio de Janeiro

Wanderson Torres



Introduction

- Since COVID-19 starts in Brazil, and a lot of people trend of moving away from big and expensive cities
- Interest in moving to Rio de Janeiro, Brazil
- Where is the safest neighborhood in Rio de Janeiro?
- What are the top neighborhoods with the most Heath Services Venues?

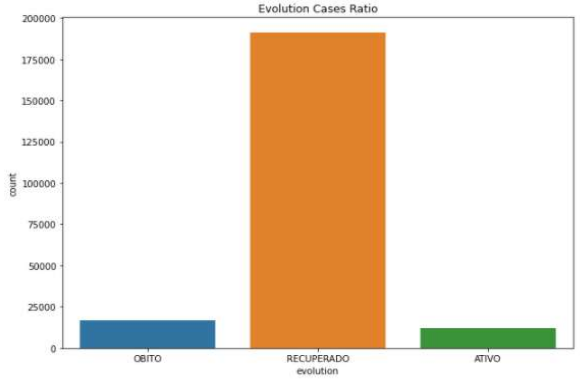
Data Sources

- Rio de Janeiro Neighborhoods : Web scrape Wikipedia + Geocoder
- Covid Dataset: RIO DATA Website
- Rio de Janeiro Population: RIO DATA Website
- Health Services Venues: Foursquare API
- Rio de Janeiro Geodata : Geojson from RIO DATA Website -> Choropleth Map with FOLIUM

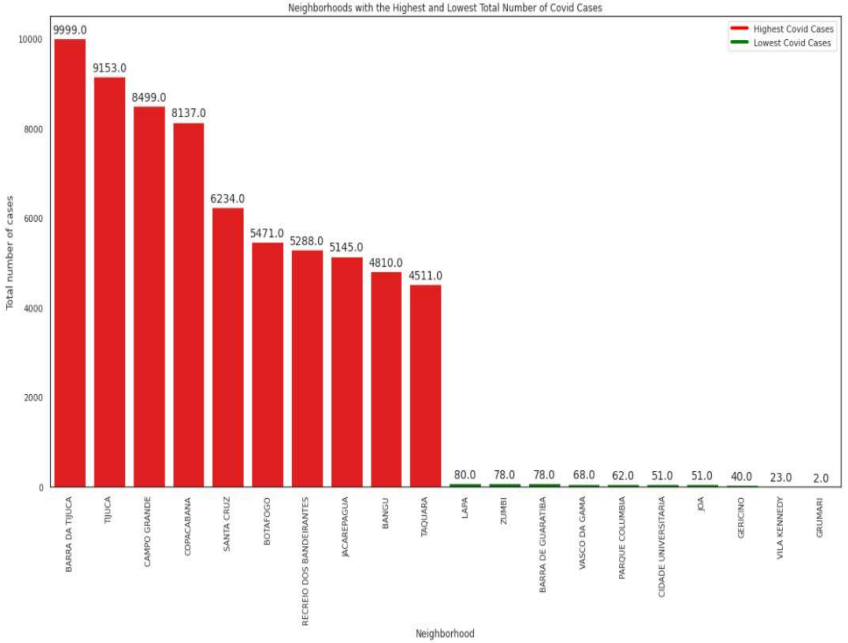
Data Cleaning into the COVID Rio Dataset

- Limited covid dataset from march, 2020 to April 15 2021
- Deleted around 20,000 records with incorrect neighborhood
- Feature Selection: Drop columns with the redundancy information
  - Sex
  - Age Group
  - Race

Exploratory Data Analysis



Neighborhood	ATIVO	OBITO	RECUPERADO	Total
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



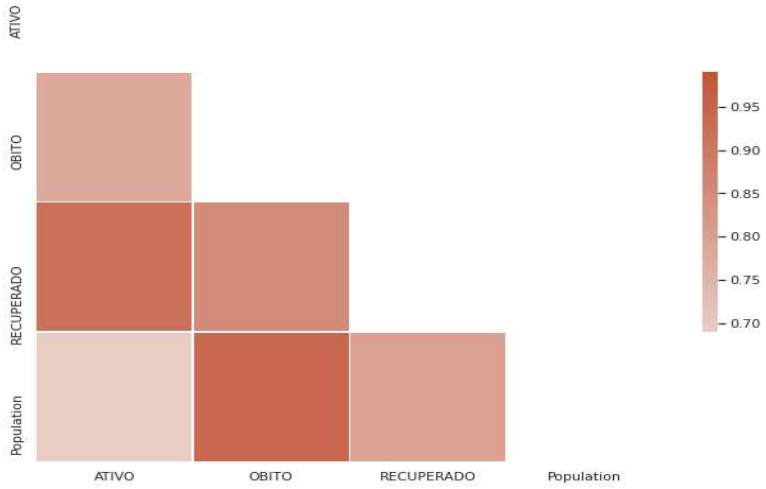
One Hot Encoding and Merge Datasets

- Transform the COVID Dataset and merge with the population

	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

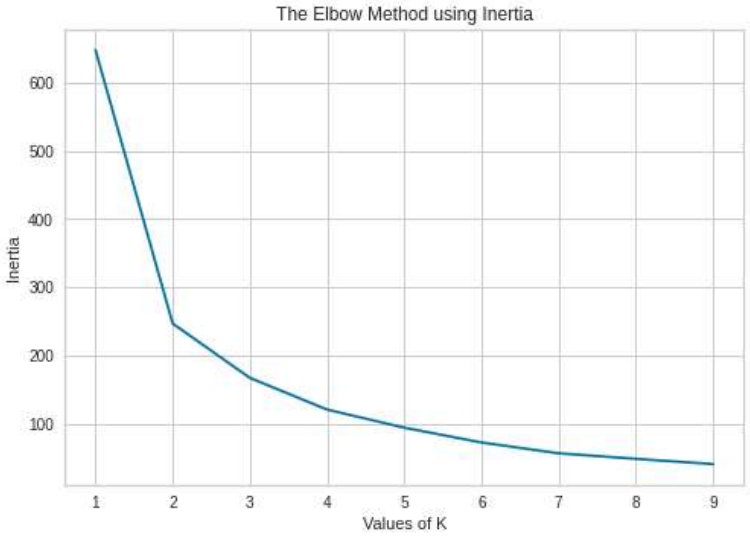
Analyzing the Correlation

	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



Cluster the Neighborhoods

- Find the best number of Clusters
  - Elbow Method:



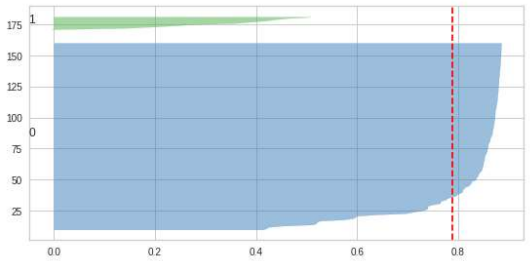
- Find the best number of Clusters
  - Silhouete Score:

- **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**

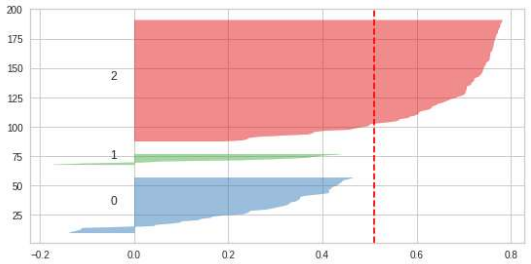
Find the best number of Clusters

- Silhouete Visualizer

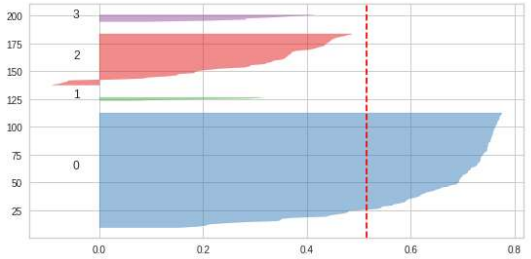
K = 2



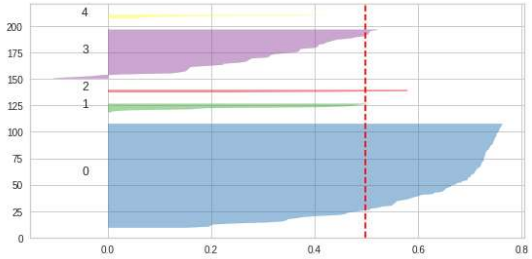
K = 3



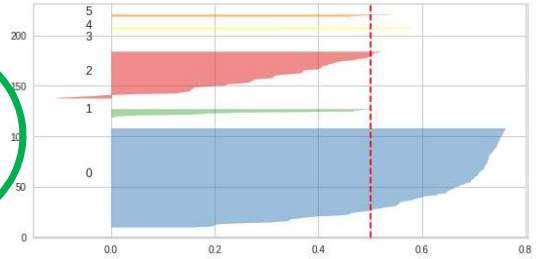
K = 4



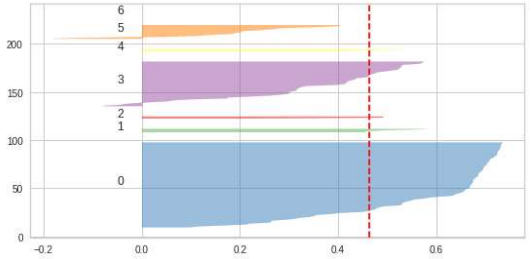
K = 5



K = 6



K = 7



K-means Clustering

- What is it and why use it? Unsupervised learning algorithm that will be used to group or clustering a large covid dataset to identify neighborhoods based on safety.
- All the preparation in the dataset for the algorithm have already done.
- Chose to group in 6 clusters.

Result from k-means clustering

Cluster 0 → green Cluster 2 → blue Cluster 4 → purple

Cluster 1 → red Cluster 3 → Orange Cluster 5 → yellow

Analyzing the Clustering

- **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

Create the COVID RISK INDEX (CRI)

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

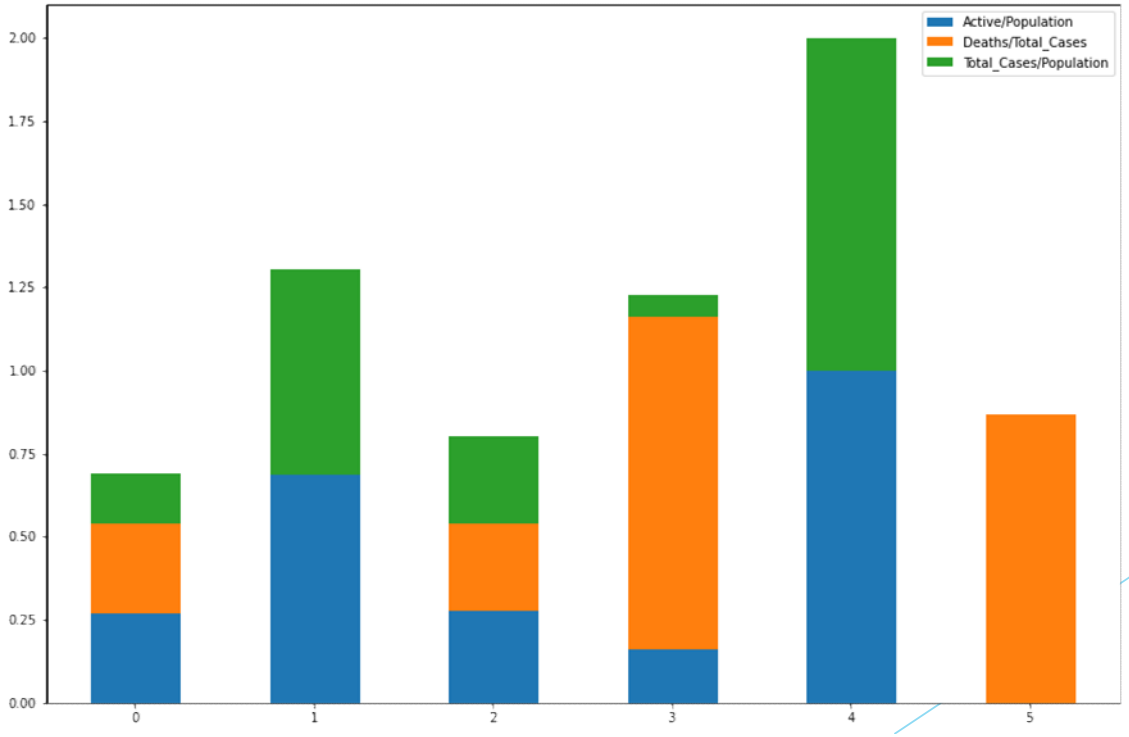
1. The number of active COVID cases by the population.
2. Number of deaths by COVID by the total number of COVID cases.
3. Total number of COVID cases by the population.

CRI = 1 + 2 + 3

COVID RISK INDEX (CRI) Analysis

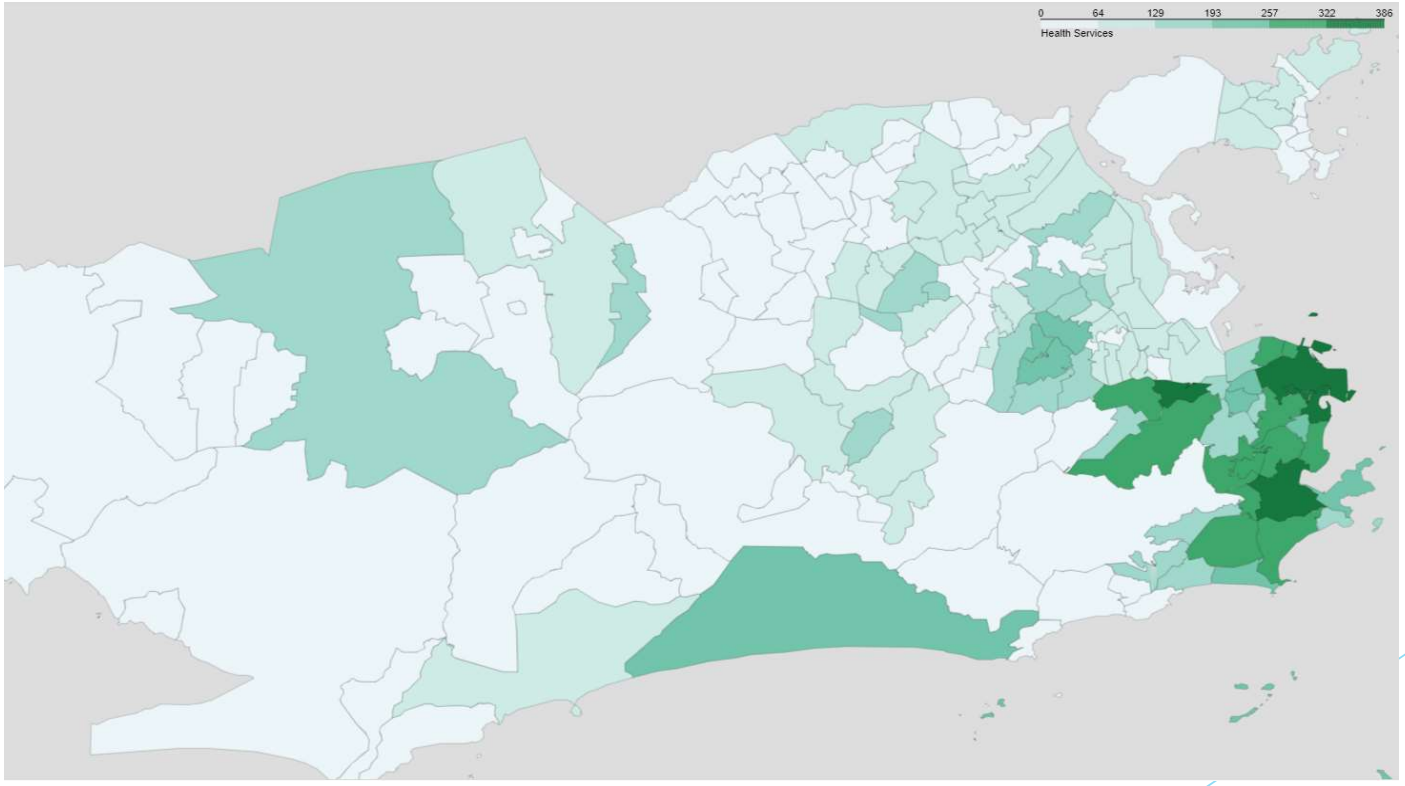
- Now, with CRI, I can better assess the risk of COVID per cluster.
- We can see the highest CRI value belongs to cluster 4 .
- That means the cluster has a High Risk to COVID.

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



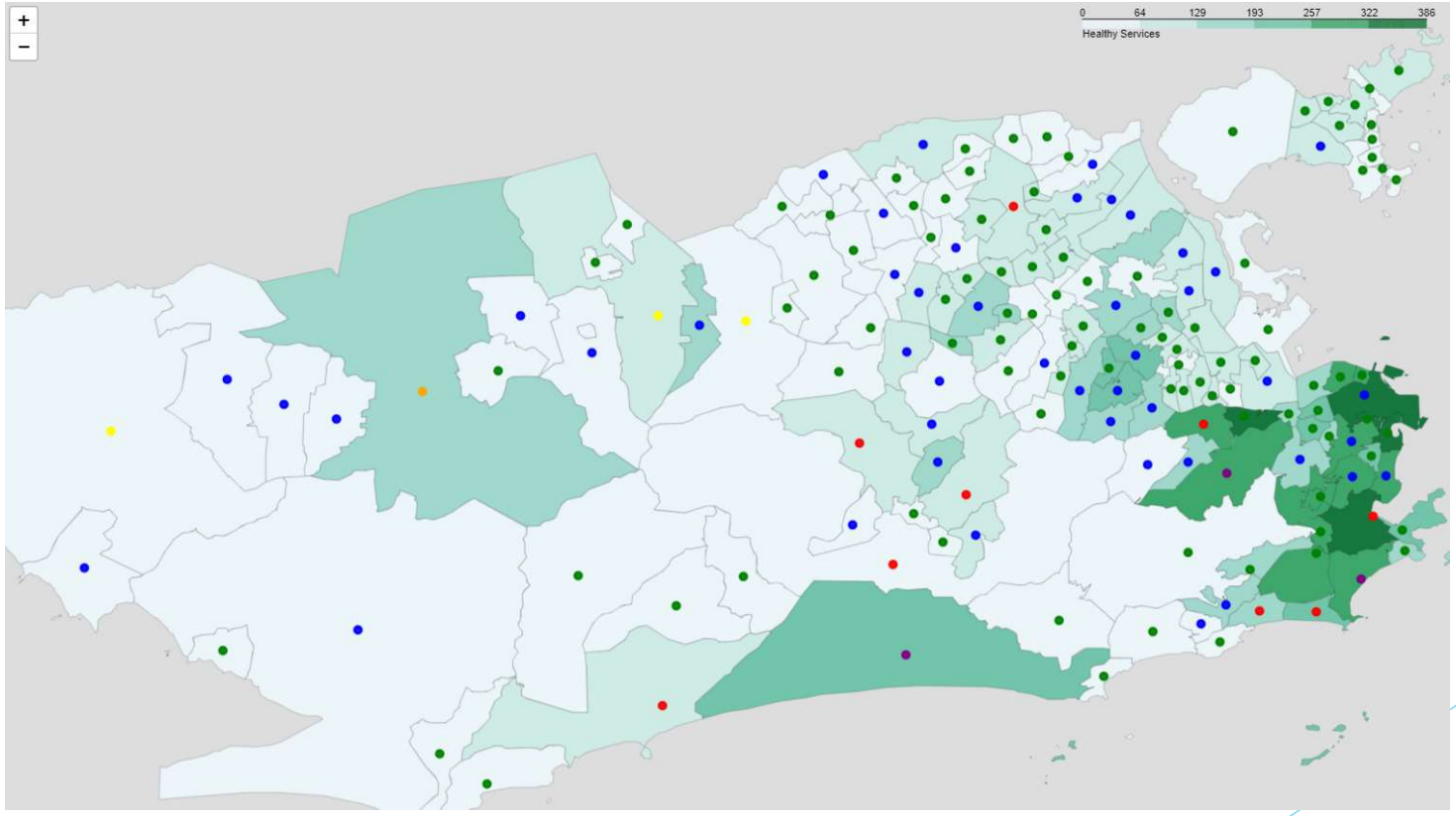
Explore the neighborhoods with health services in Rio de Janeiro using Foursquare API

- Create the Choropleth Map using ***FOLIUM python library*** with the number of health services points like hospitals, clinics, drugstores, medical offices,…



Choose a safer place to live

- Combining the COVID risk clustering and the health service offerings

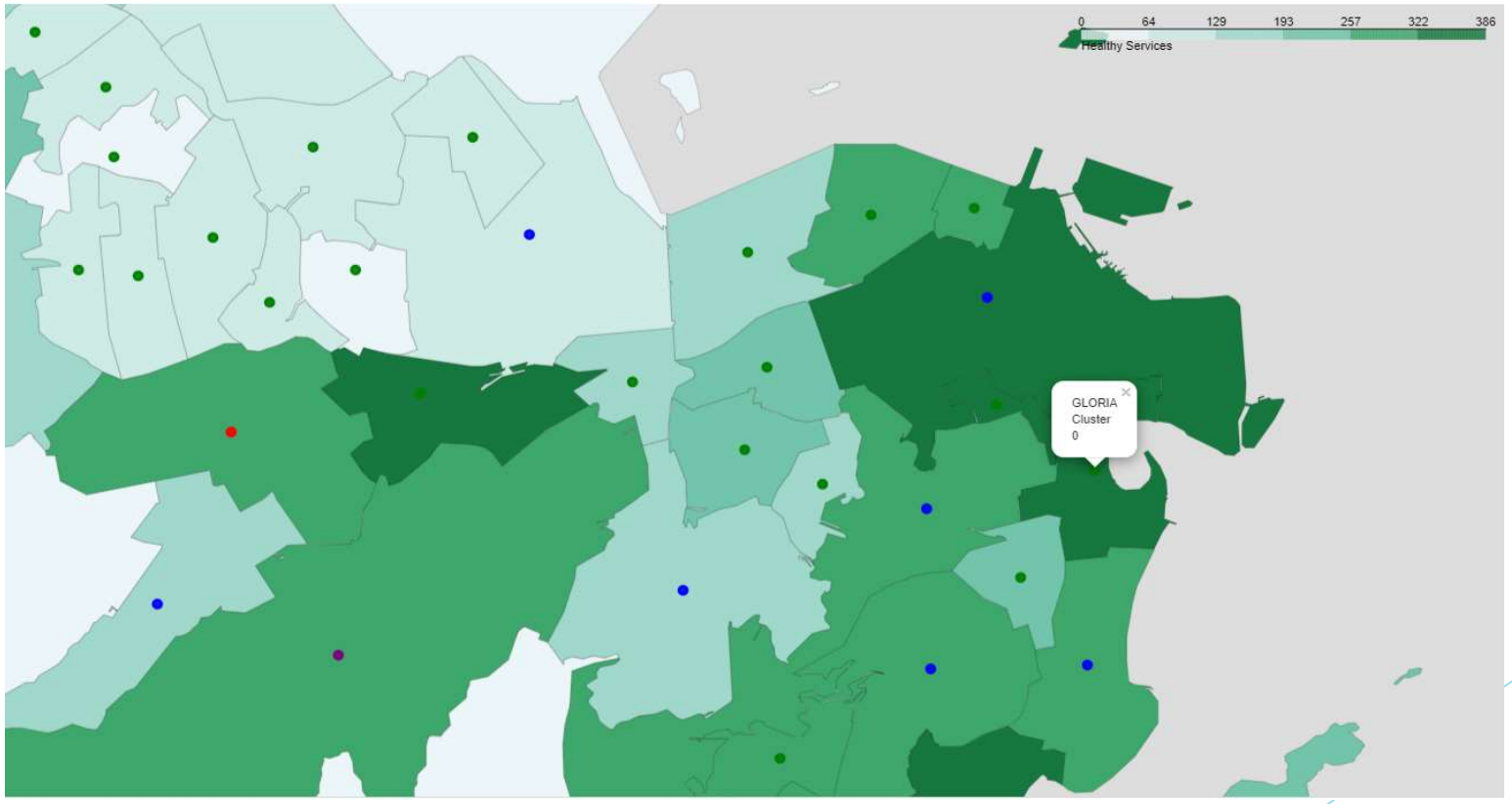


Cluster	Color	CRI
Cluster 0	green	0.69
Cluster 2	blue	0.80
Cluster 5	yellow	0.87
Cluster 3	orange	1.23
Cluster 1	red	1.30
Cluster 4	purple	2.00

+ RISK

Choose a safer place to live

- Getting deeper in the dark green areas



Summary / Conclusion

- Six clusters were identified by K-means algorithm.
- The Covid risk index (CRI) were created to compare the clusters.
- Safest Cluster to live in: Cluster 0
- 2 best options of Neighborhoods considering the CRI and Heath Services offerings: LAPA and GLÓRIA
- Discussions points:
  - Subnotifictions in poor areas
  - Foursquare is not very famous in Brazil
- Futures Analysis:
  - Include the sex, age group and race features of COVID cases x Neighborhoods
  - Use others algorithms and compare the results