State High Schools in Rio de Janeiro (Brazil)

Capstone Final Project

Daniela Verbicário Botelho da Costa

July 25, 2019

Index

1 Introduction

2 Data Collection

3 Methodology

4 Analysis

5 Results and Discussion

6 Conclusion

Introduction

1
2
3
4
5
6

⇨ High schools are responsibility of the states in Brazil.

⇨ 80% of all high school offers in the state of Rio de Janeiro provided by the state.

⇨ 20.000 students did not have their places in 2019.

⇨ Immediate need to increase the offer of public high schools.

Question to be answered

⇨ Which neighborhoods could beneficiate more from those new public high schools in the city of Rio de Janeiro ?

Stakeholders

⇨ Secretary of Education of Rio

⇨ City citizens

Data Collection

1
2
3
4
5
6

Data.Rio

- Geographic limits of each neighborhood in the city of Rio

Foursquare and Escol.as

- High schools in the city of Rio

Brazilian Institute of Geography and Statistics

- Population per neighborhood (2010)
- HDI-Income per neighborhood (2010)

Methodology

1
2
3
4
5
6
8

1. Collection of data from the city of Rio to know precisely the identification and coordinates of each neighborhood;

2. Using Foursquare API for retrieving more schools in the city of Rio de Janeiro;

3. Scraping high schools in Rio de Janeiro from escol.as;

4. Selecting only state high schools from all the schools retrieved;

5. Aggregating schools per neighborhood;

6. Importing dataframe with socioeconomic features from local disc;

7. Exploring the data: Visualization of features in a scatter plot;

8. Clustering algorithm: K-means;

9. Creating cluster label column in the main dataframe and examine clusters;

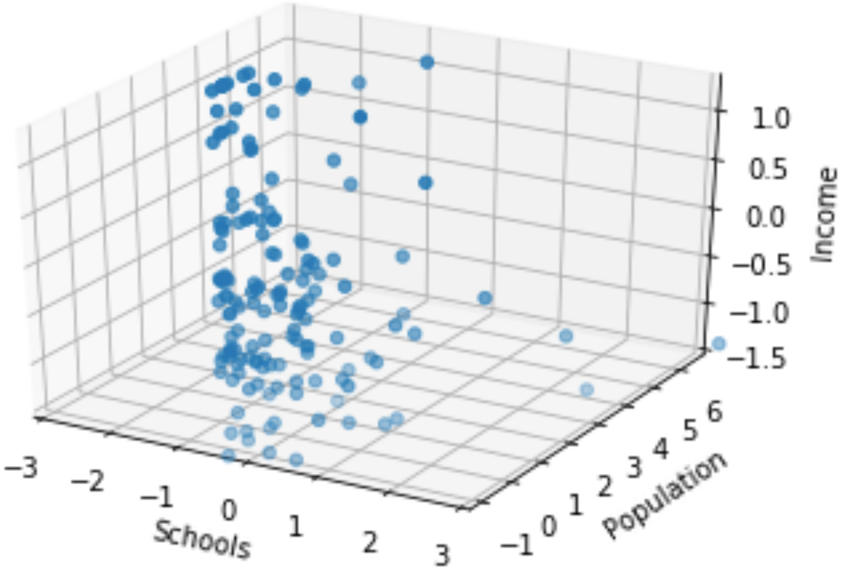
10. Classifying each cluster according to its characteristics;

11. Applying criterion of density per school to make the final choice;

12. Mapping chosen neighborhoods.

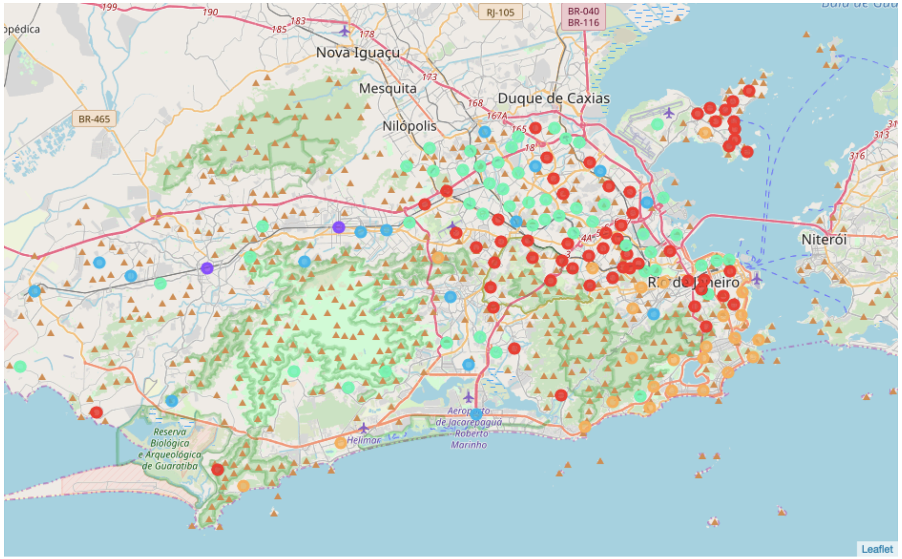
Analysis

1
2
3
4
5
6



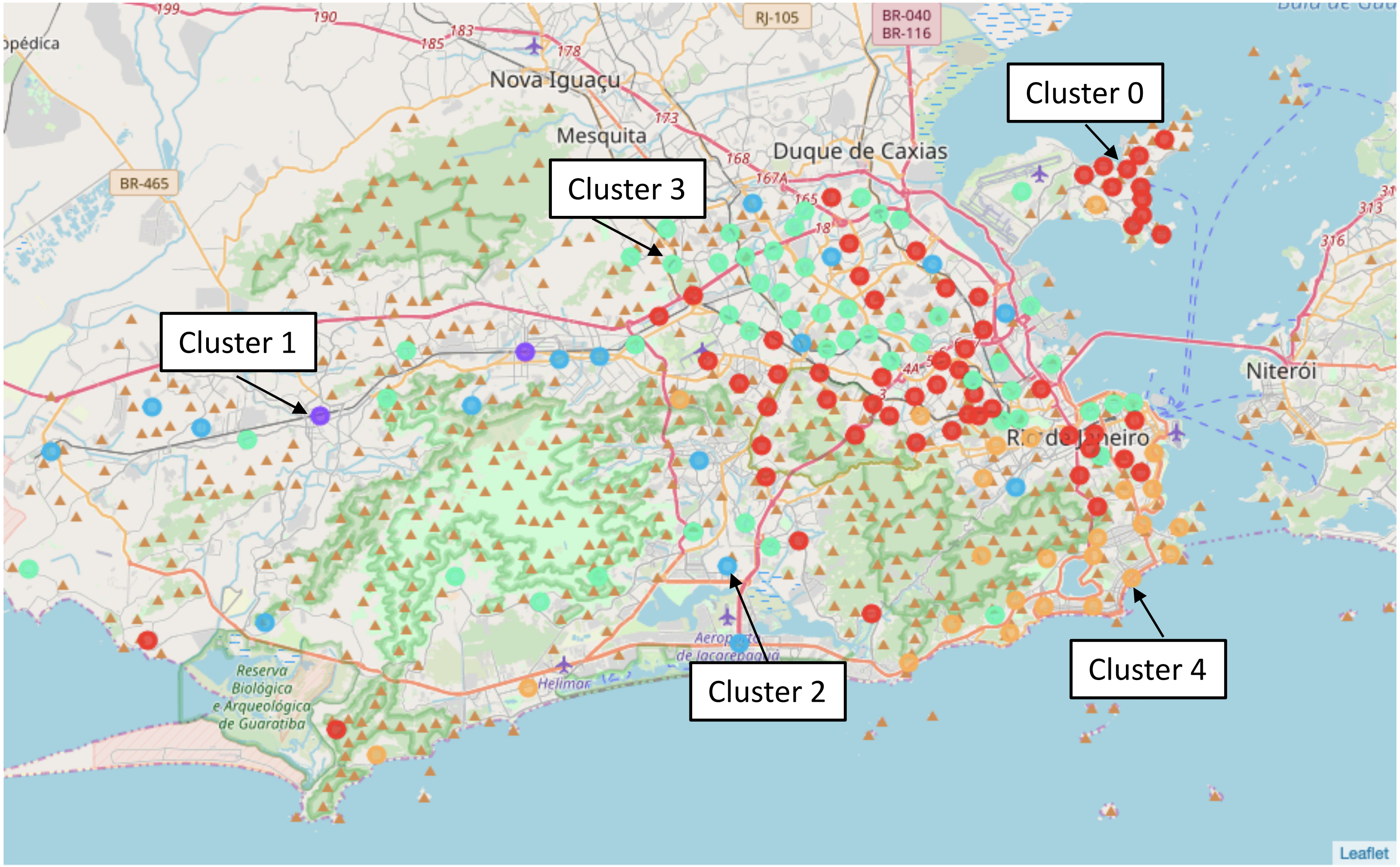
K-means Clustering

5 clusters



Analysis

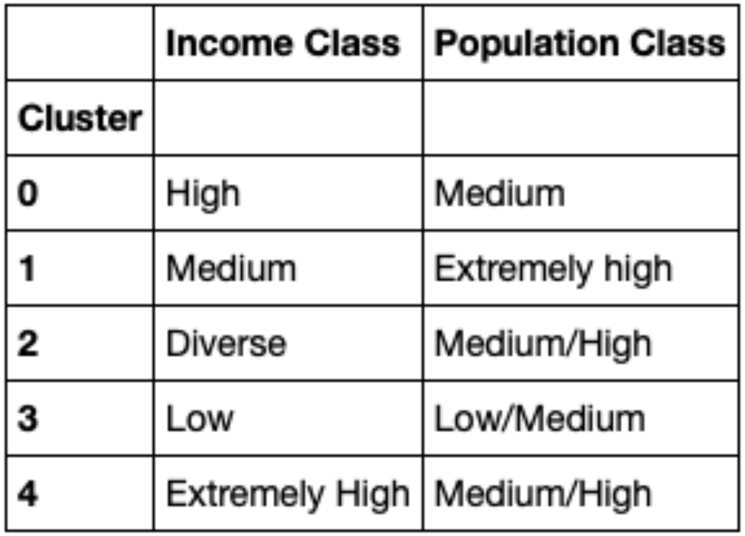
1
2
3
4
5
6



Analysis

1
2
3
4
5
6

⇨ Classification based on income and population



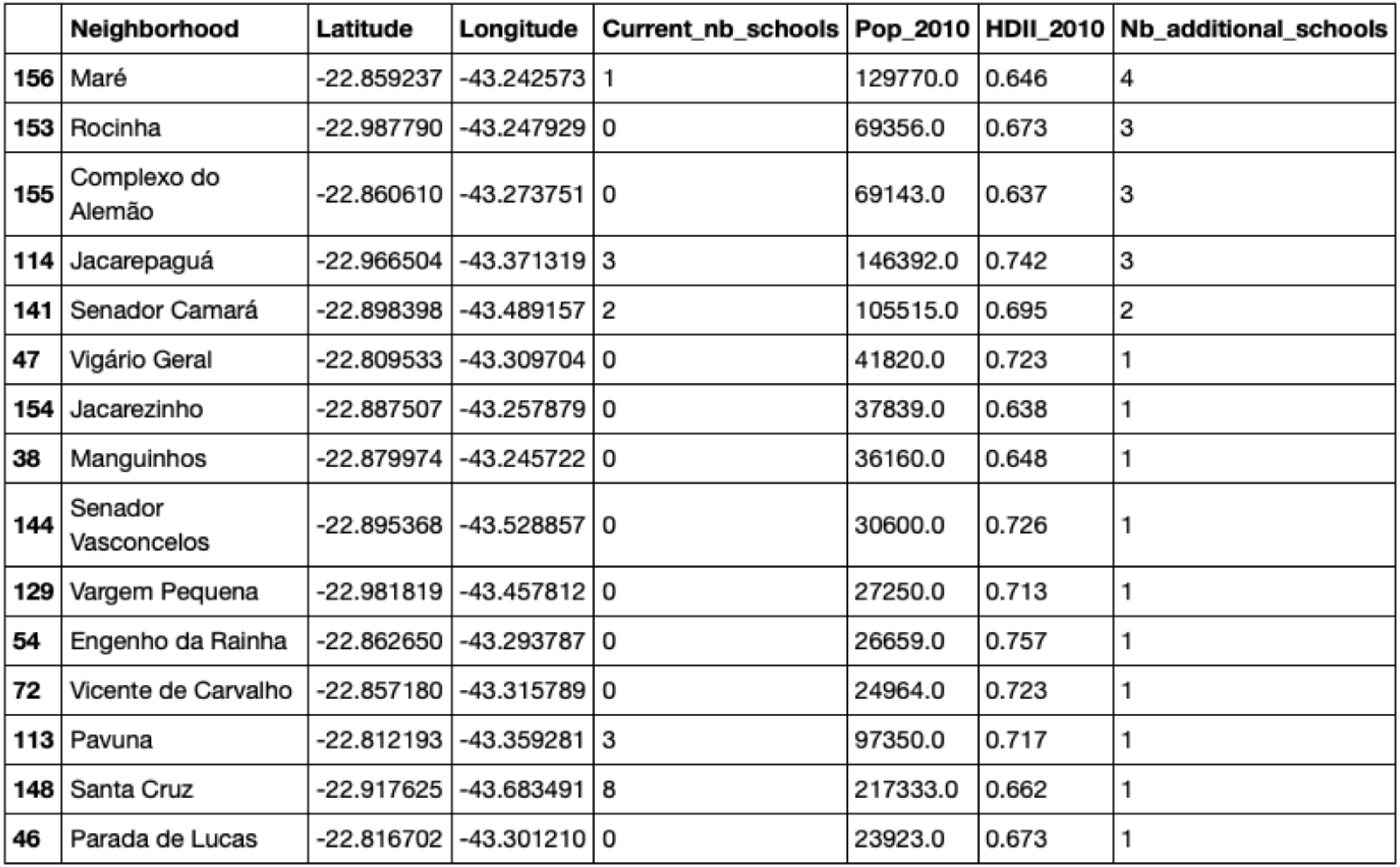
	Income Class	Population Class
Cluster		
0	High	Medium
1	Medium	Extremely high
2	Diverse	Medium/High
3	Low	Low/Medium
4	Extremely High	Medium/High

⇨ Selected clusters: 1, 2 (income > 0.750) and 3

⇨ For each resulting neighborhood, calculated the number of people per school in case of one, two, three, four and five more schools and saved these numbers in a list

⇨ The 25 highest numbers were identified along with their neighborhoods to produce the final table

Results and Discussion



	Neighborhood	Latitude	Longitude	Current_nb_schools	Pop_2010	HDII_2010	Nb_additional_schools
156	Maré	-22.859237	-43.242573	1	129770.0	0.646	4
153	Rocinha	-22.987790	-43.247929	0	69356.0	0.673	3
155	Complexo do Alemão	-22.860610	-43.273751	0	69143.0	0.637	3
114	Jacarepaguá	-22.966504	-43.371319	3	146392.0	0.742	3
141	Senador Camará	-22.898398	-43.489157	2	105515.0	0.695	2
47	Vigário Geral	-22.809533	-43.309704	0	41820.0	0.723	1
154	Jacarezinho	-22.887507	-43.257879	0	37839.0	0.638	1
38	Manguinhos	-22.879974	-43.245722	0	36160.0	0.648	1
144	Senador Vasconcelos	-22.895368	-43.528857	0	30600.0	0.726	1
129	Vargem Pequena	-22.981819	-43.457812	0	27250.0	0.713	1
54	Engenho da Rainha	-22.862650	-43.293787	0	26659.0	0.757	1
72	Vicente de Carvalho	-22.857180	-43.315789	0	24964.0	0.723	1
113	Pavuna	-22.812193	-43.359281	3	97350.0	0.717	1
148	Santa Cruz	-22.917625	-43.683491	8	217333.0	0.662	1
46	Parada de Lucas	-22.816702	-43.301210	0	23923.0	0.673	1

Results and Discussion

1
2
3
4
5
6

➯ Most vulnerable areas are important *favelas*.

Coherent since they are densely populated and have a low level of income

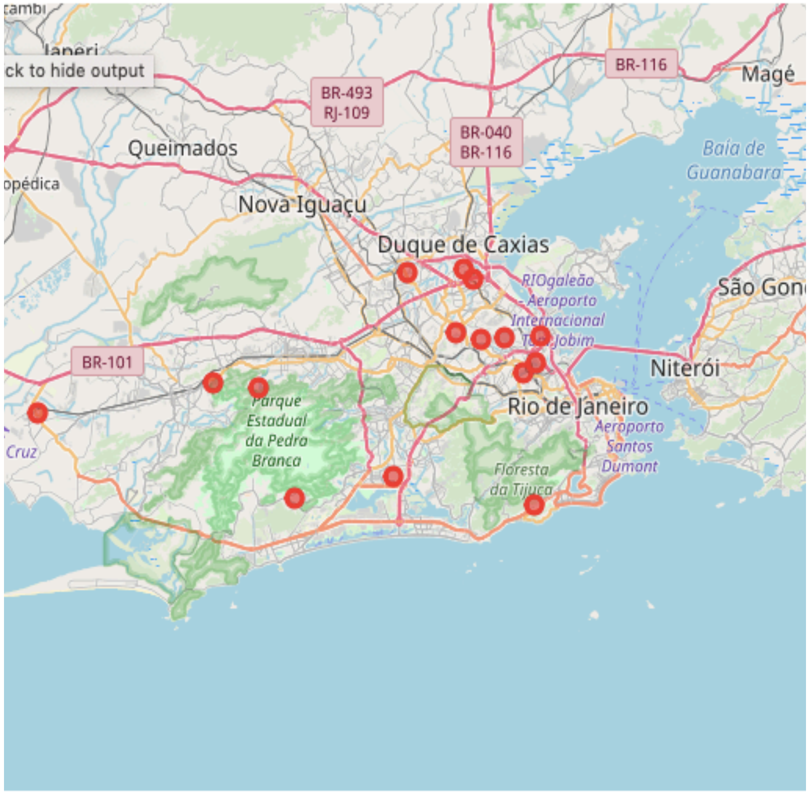
➯ This approach have some limitations:

(i) Most of data dates back to 2010;

(ii) The population as a proxy for the demand can be misleading if the pyramid of ages of neighborhoods differ significantly;

(iii) Only three factors were considered and one kind of clustering algorithm tested.

➯ Recommended zones should therefore be considered only as a starting point for more detailed analysis.



Conclusion

The analysis took into consideration the current number of public high schools, the population and the level of income of each neighborhood as proxys of the offer and the demand.

Clustering of neighborhoods was then performed in order to create major zones of interest used as starting points for final exploration.

Final criterion was the density per school in the case of the existence of more schools in each preselected area. Those with the highest demand in populational terms were then chosen.

14 vulnerable neighborhoods were chosen as priorities. For some of them, more than one school would be needed to balance the demand and the offer.