INTRODUCTION

This project was designed to serve as a final assessment for the Data Science certification of coursera, IBM's online course, as well as to serve as an intermediary assessment of the Data Science Infnet course, Data Analytics and Machine Learning. The objective of the project is to answer the question: where's the best location to open the "FastSnack" cafeteria? It is important to note that the scope of the project limited the search for the ideal location to the surroundings of the Botafogo neighborhood, located in the city of Rio de Janeiro, in Brazil.

Botafogo is a neighborhood located in the prime area of one of the main cities in Latin America, which makes it an extremely attractive place, however it imposes some difficulties. The high HDI of 0.952 in the region suggests a good financial condition for its residents, in addition the neighborhood has a strong commercial area, which generates a large volume of circulation of people, both factors are attractive for the establishment of businesses. However, the reasons that drive the interests of establishing the cafeteria also attract several other competitors.

Two niches of diner customers are very present in Botafogo, workers and students, both constantly in search of fast and practical food to consume between their chores. FastSnack opted for the area precisely to benefit from both niches, however the focus will be directed to serving students, due to the easy location of their concentration points as well as the low demand regarding the food consumed. Once this direction of the business is defined, the focus to be adopted to achieve the project objective is clear, seeking the place with the highest concentration of schools.

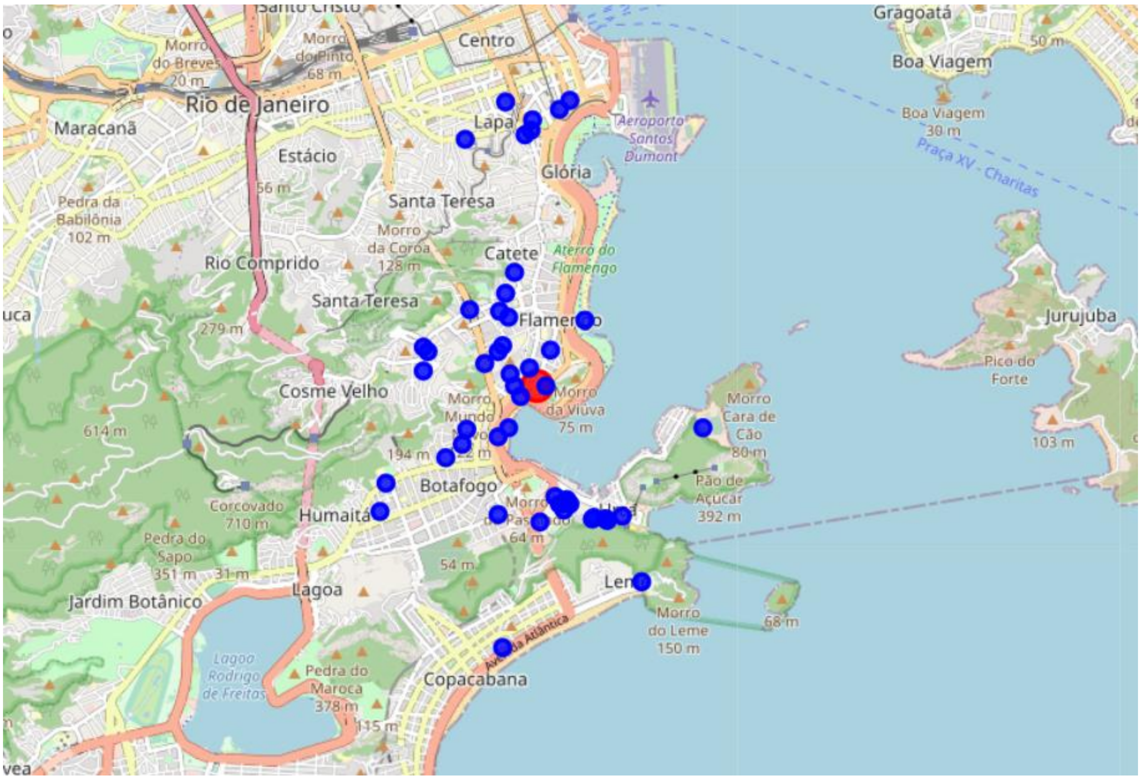
However, to better support the choice, the number of competitors in the surroundings will also be considered. This search was limited to a small radius due to the nature of the business being aimed at the convenience of those who are close. Thus, the aim of the project is to find a location in the designated region with a balance between a high number of schools and a small number of competitors. This choice was based on data science tools and techniques, following its methodology in order to find the point that balances the two variables mentioned.

The database supporting the entire analysis was Foursuare, which offers a variety of reliable information. After data collection, pre-processing, cleaning, filtering and formatting were performed in order to make them suitable for use in the model. The predefined scope of using a clustering machine learning, specifically K-means is adequate, since the need for the problem is to identify regions where there were a high number of schools. Finally, the analysis of the number of competitors takes place in the clusters generated through the concentration of schools.

DATA ACQUISITION

The scope of the course defines that the Foursquare API must be used to acquire the data. Although it is allowed to use other data sources simultaneously, the process was not necessary. The Foursquare API has all the data raised as necessary to meet the objective. These data are the number of schools and restaurants in selected regions. Lightning was established. To increase the number of possible daily requests, the credentials of a developer account were used.

The project limited its search area to the center of Botafogo and its surroundings. A radius of 3 kilometers was drawn from the address Praia de Botafogo number 328.



The blue dots point to the locations of the schools while the red dot points to the central point from which the radius for the search was measured. In a brief visual analysis, it is already plausible to identify a possible cluster to the north of the image, separated from the rest of the schools by an empty strip.

DISCUSSION OF RESULTS AND CONCLUSION

An interesting question to be answered is how much the added competition in the analysis influences the choice of the best place to open FastSnack. The following table represents the distribution of schools in each cluster.

Cluster number	Number of schools	Percentage of schools
0	7	14,3%
1	25	51,0%
2	17	34,7%

Considering only the variable number of schools, it is clear which region we should choose. Cluster 1 has approximately half of the total schools analyzed, so in this first analysis FastSnack's ideal location would be in the cluster 1 center, the region where it would be able to serve its 25 schools in the best way. In addition, a factor that had not been expected is that cluster 1 is in the center of Botafogo, although the proximity of the center of the neighborhood was not a variable that was intended to be achieved in the end turns out to be a satisfactory coincidence. In contrast, cluster 0 has a low concentration of schools, 7 in total. Cluster 2 has approximately 35% of schools, a lower number compared to cluster number 1, however it still has a good volume of schools in its surroundings.

Analyzing the number of competing restaurants in each cluster we have:

Cluster number	Number of restaurants	Percentage of restaurants
0	50	70,4%
1	15	21,1%
2	6	8,5%

Again, cluster 0 was in the worst position with more competitors nearby. However, in this analysis the best placed was cluster 2 with two and a half times less competitors than cluster 1. However, the main variable since the beginning of the project was the number of schools in the vicinity and this difference found in the proximity to competitors it is not big enough to change the decision to an ideal location. So the best place to open FastSnack is in the center of cluster 1.