**Identifying best locations to open a business**

Rafael de Sant’Anna Chagas

February 04, 2021

1. **Introduction**

Entrepreneurs have always struggled when trying to find the best location to open a business, and many don’t succeed. In order to make finding the right spot to set up a specific business easier for entrepreneurs all around the world, thus preventing unnecessary failures, we can use data to compare various locations and identify patterns that will lead us to the places where the business in question will have the highest probabilities of success based on similar, but functional, businesses.

A python application was developed to fulfil this purpose. The application’s algorithm, logic and methodology will be discussed throughout this report.

1. **Data**

There has been used two datasets. The first and most important one is provided by Foursquare and it lists and describes the most relevant venues, up to a hundred, in a given area. There is a great deal of information in the descriptions provided by Foursquare, however only the venues ids, coordinates and categories has been used in these analyses. The other dataset is provided by Nominatim and contains relationships between addresses and coordinates, all of which has been used in these analyses.

Figure 1 - Foursquare logo Figure 2 - Nominatim logo

1. **Methodology**

For the purpose of analyzing the application and its results, this report will take the city of São Paulo, located in Brazil, as an example. As the chosen venue category, “Theater” will be used.

* 1. **Mapping region areas**

After a desired city to open the venue has been specified, the application divides the city in 1000-meter radius areas. In order to accomplish that, the application starts scanning all the coordinates in a circular pattern from the center of the city outwards, increasing the scanned perimeter, all the while getting the cities of all coordinates calculated by using the Nominatim API. When all coordinates from the outer circle return cities that are different from the chosen city, the coordinates computation stops. Then, all the coordinates that returned a city that is different from the chosen one, are deleted, leaving the application only with the coordinates pertaining to the chosen city.

It’s worth it to mention that all the areas’ centers are offset about 1500 meters from one another.

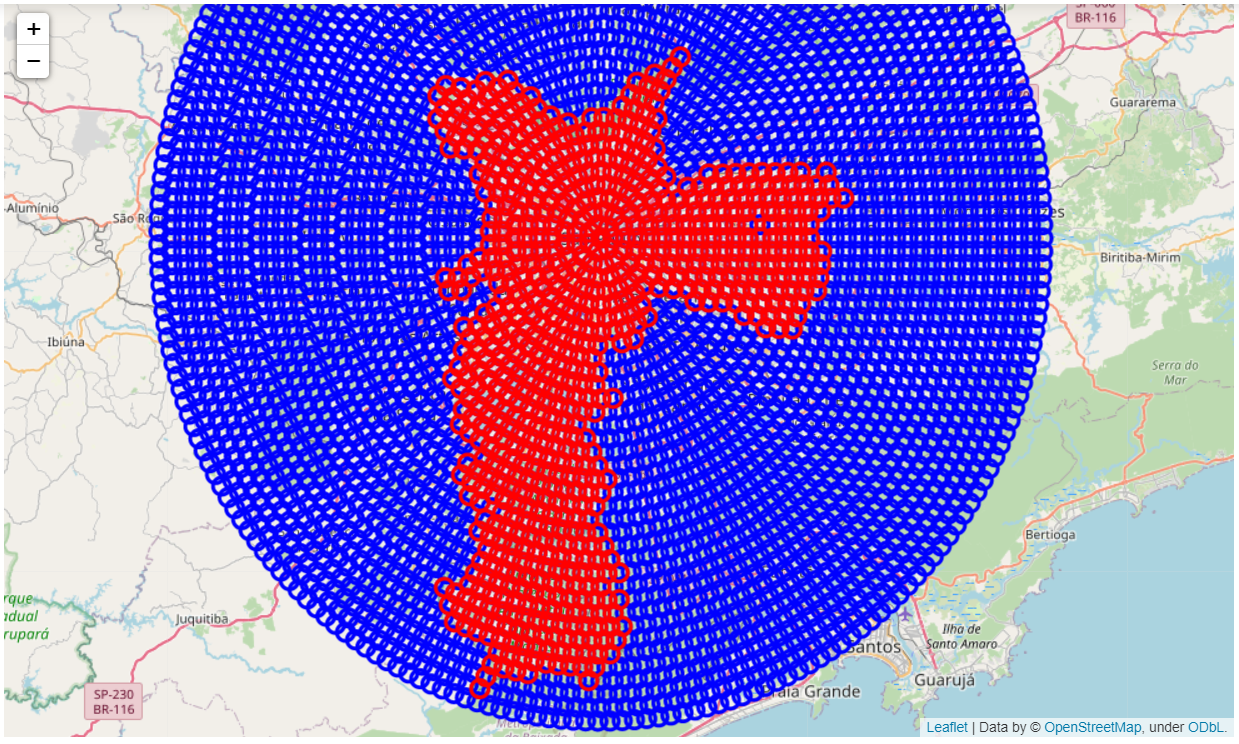


Figure 3 – City – São Paulo (red area)

* 1. **Fetching city venues**

The next step is fetching the hundred most important venues and their details of each remaining locations. This is achieved by using the Foursquare API.

After getting all the data, the application will extract the venues’ ids, categories and coordinates, cleaning the data.

* 1. **Dealing with venues duplicates**

Because the Foursquare API only allows the user to fetch venues by specifying a location and a range radius, the application has to overlap all areas within the city to maximize its range. However, when doing so, lots of venues are computed more than once. In order to fix this problem, all recuring venues are attributed to the nearest mapped area coordinate.

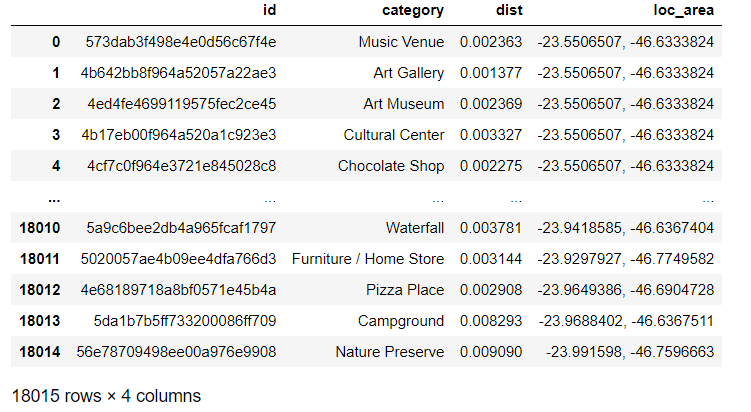


Figure 4 - Venues

* 1. **Creating locations’ patterns vectors**

Each location is analyzed according to its venues’ categories’ count. Then, these counts are used to create locations’ vectors. These vectors contain the number of venues in the category for each category.

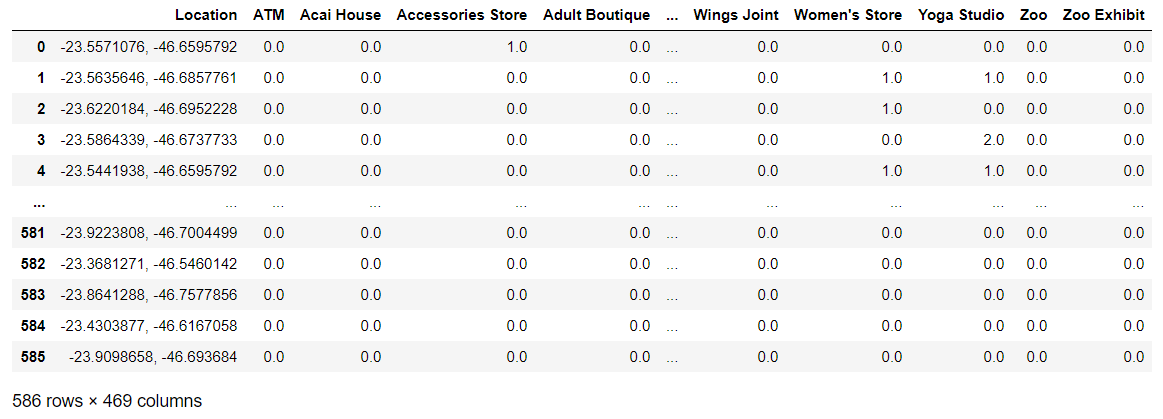


Figure 5 - locations' vectors

* 1. **Grouping city areas into clusters**

In order to understand the most similar regions, the application uses a K-means classifier, an unsupervised machine learning algorithm.

In the first part, the application user must choose the best quantity of clusters. The algorithm iterates through numbers of clusters until it reaches 100 or the number of available areas in the city. After this process, it provides the user a scatter plot containing the mean distances from samples to their respective cluster centers. In this example the number of clusters was chosen by the variations of the variation rates between means distances. Thus, 9 was the clusters quantity chosen.

In the second part, the application groups all city areas into the different, previous chosen, clusters.

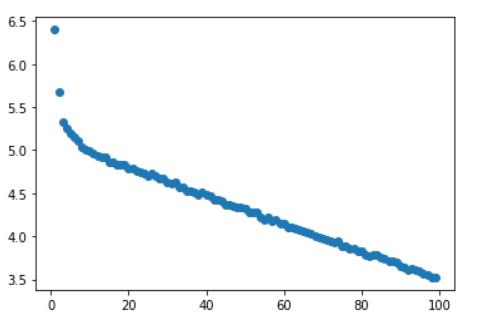


Figure 6 - variations of the variation rates between means distances

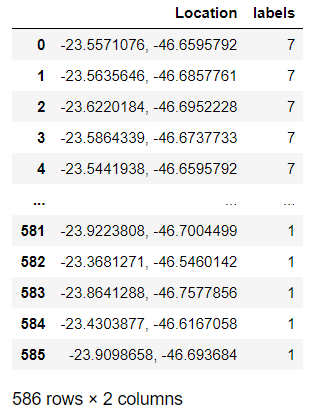


Figure 7 - Clusters division

* 1. **Identifying the best location**

The application calculates the mean occurrence for each venue category in each cluster. Then, it points out the cluster where the chosen venue category mean occurrence is higher, cluster 7 in this case. After that, the application finds the best areas to open the venue by identifying the regions, within that cluster, where the chosen venue category is lower than the cluster weighted mean.

The best locations to open the desired business are listed on figure 9 because they present the pattern where this kind of business usually thrives or, at least, apparently survives, and these locations also present the lowest “Theater” occurrence in the cluster.

Figure 10 represents locations that present the best opportunities among the best locations within the best cluster.

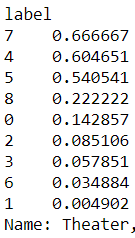
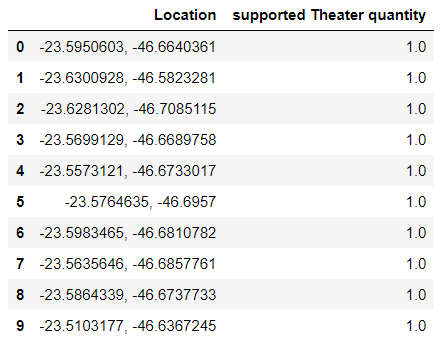
 

Figure 8 - Theater mean occurrence Figure 9 - Best locations

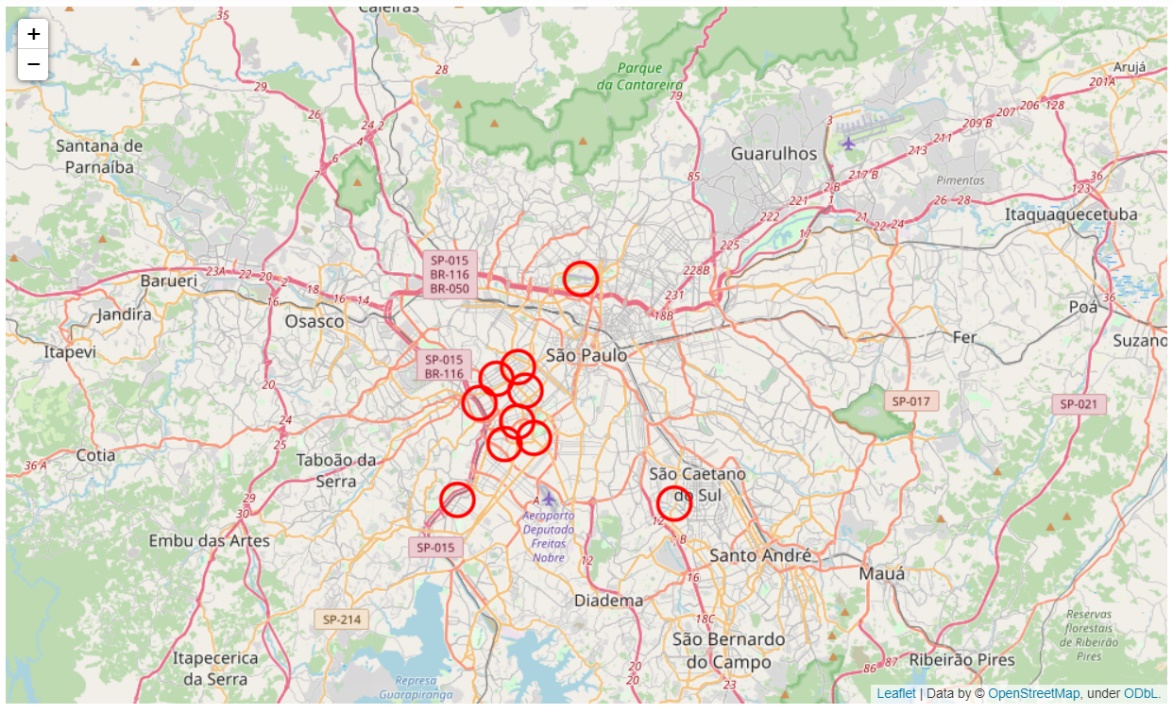


Figure 10 - Map of the best locations (red: best opportunities)

1. **Results**

The application operates almost independently, and it gives the user a good location recommendation to open a chosen business if that’s the case. Nevertheless, it will tell the user whether the city they chose can’t accommodate their business. The use of machine learning algorithms were essential to the application’s success. However, it may demand a rather powerful connection to the internet to perform the required tasks this application executes.

1. **Discussion**

Despite the fact that the application provides good sites recommendations, it lacks information that could only be gathered by human experience, sometimes, intuitively. Furthermore, the chosen venue category might not be present in the desired city or have a low occurrence, which impairs the application’s efforts.

1. **Conclusion**

Nowadays, it is of the utmost importance that entrepreneurs make wise choices, especially when opening a new business. Large amounts of money are wasted because they usually don’t take data into account. However, this application offers every entrepreneur the necessary support to choose the right spot for their business. Providing the best options of locations that their businesses require, this application shows itself useful beyond expectations.