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.

2. 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

3. 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.

3.1. 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.

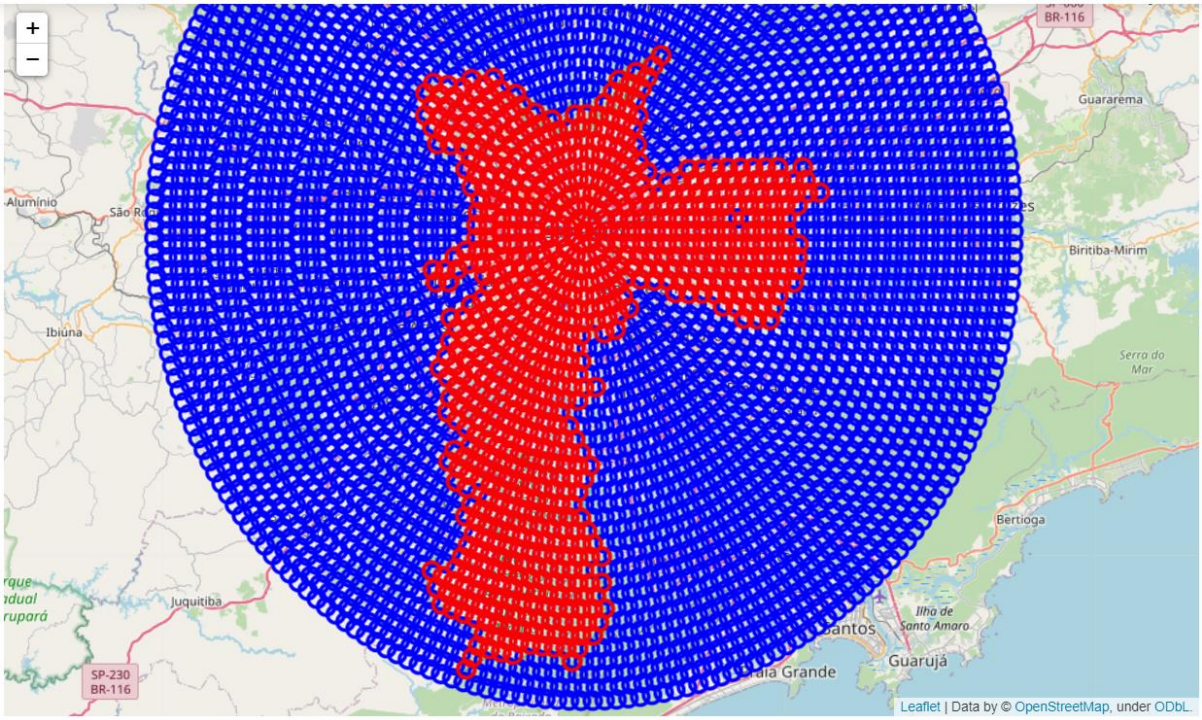

Figure 3 – City – São Paulo (red area)

3.2. 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.

3.3. 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.

	id	category	dist	loc_area
0	573dab3f498e4e0d56c67f4e	Music Venue	0.002363	-23.5506507, -46.6333824
1	4b642bb8f964a52057a22ae3	Art Gallery	0.001377	-23.5506507, -46.6333824
2	4ed4fe4699119575fec2ce45	Art Museum	0.002369	-23.5506507, -46.6333824
3	4b17eb00f964a520a1c923e3	Cultural Center	0.003327	-23.5506507, -46.6333824
4	4cf7c0f964e3721e845028c8	Chocolate Shop	0.002275	-23.5506507, -46.6333824
...	...	...	...	...
18010	5a9c6bee2db4a965fcaf1797	Waterfall	0.003781	-23.9418585, -46.6367404
18011	5020057ae4b09ee4dfa766d3	Furniture / Home Store	0.003144	-23.9297927, -46.7749582
18012	4e68189718a8bf0571e45b4a	Pizza Place	0.002908	-23.9649386, -46.6904728
18013	5da1b7b5ff733200086ff709	Campground	0.008293	-23.9688402, -46.6367511
18014	56e78709498ee00a976e9908	Nature Preserve	0.009090	-23.991598, -46.7596663

18015 rows × 4 columns

Figure 4 - Venues

3.4. 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.

	Location	ATM	Acai House	Accessories Store	Adult Boutique	...	Wings Joint	Women's Store	Yoga Studio	Zoo	Zoo Exhibit
0	-23.5571076, -46.6595792	0.0	0.0	1.0	0.0	...	0.0	0.0	0.0	0.0	0.0
1	-23.5635646, -46.6857761	0.0	0.0	0.0	0.0	...	0.0	1.0	1.0	0.0	0.0
2	-23.6220184, -46.6952228	0.0	0.0	0.0	0.0	...	0.0	1.0	0.0	0.0	0.0
3	-23.5864339, -46.6737733	0.0	0.0	0.0	0.0	...	0.0	0.0	2.0	0.0	0.0
4	-23.5441938, -46.6595792	0.0	0.0	0.0	0.0	...	0.0	1.0	1.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...
581	-23.9223808, -46.7004499	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
582	-23.3681271, -46.5460142	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
583	-23.8641288, -46.7577856	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
584	-23.4303877, -46.6167058	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
585	-23.9098658, -46.693684	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0

586 rows × 469 columns

Figure 5 - locations' vectors

3.5. 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.

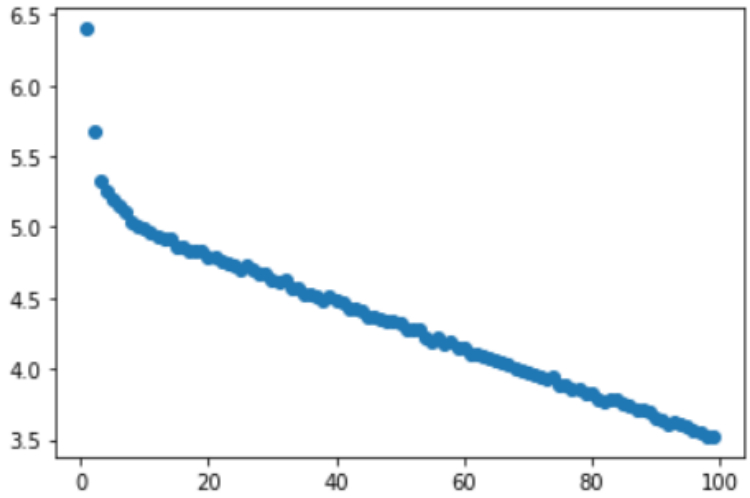

Figure 6 - variations of the variation rates between means distances

	Location	labels
0	-23.5571076, -46.6595792	7
1	-23.5635646, -46.6857761	7
2	-23.6220184, -46.6952228	7
3	-23.5864339, -46.6737733	7
4	-23.5441938, -46.6595792	7
...	...	...
581	-23.9223808, -46.7004499	1
582	-23.3681271, -46.5460142	1
583	-23.8641288, -46.7577856	1
584	-23.4303877, -46.6167058	1
585	-23.9098658, -46.693684	1

586 rows × 2 columns

Figure 7 - Clusters division

3.6. 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.

```
label
7    0.666667
4    0.604651
5    0.540541
8    0.222222
0    0.142857
2    0.085106
3    0.057851
6    0.034884
1    0.004902
Name: Theater,
```

Figure 8 - Theater mean occurrence

	Location	supported Theater quantity
0	-23.5950603, -46.6640361	1.0
1	-23.6300928, -46.5823281	1.0
2	-23.6281302, -46.7085115	1.0
3	-23.5699129, -46.6689758	1.0
4	-23.5573121, -46.6733017	1.0
5	-23.5764635, -46.6957	1.0
6	-23.5983465, -46.6810782	1.0
7	-23.5635646, -46.6857761	1.0
8	-23.5864339, -46.6737733	1.0
9	-23.5103177, -46.6367245	1.0

Figure 9 - Best locations

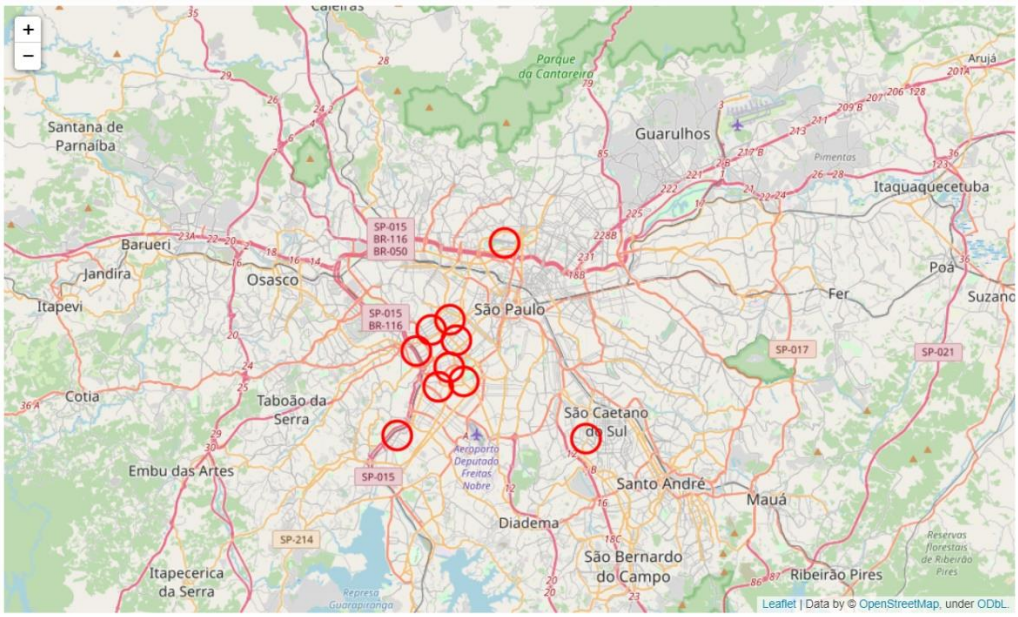

Figure 10 - Map of the best locations (red: best opportunities)

4. 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.

5. 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.

6. 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.