Determining the best location to open a Tech Startup in Florianópolis (Brazil)

Rômulo Fernandes Evangelista

May 10, 2020

1. Introduction

1.1 Background

Florianópolis is the capital city of the state of Santa Catarina, in the south of Brazil. Traditionally known for its beautiful beaches and tourist attractions, this city has seen a rise in entrepreneurship as the state and city governments issue tax incentives towards startups and new companies. Additionally, this charming city has some of the country's top universities, full of students eager to either start the next Brazilian unicorn, or help said unicorn to achieve its full potential.

These changes have attracted many dedicated and competent incubators, which now accumulate many stories of success on their portfolio.

The city, however, extends itself over an area of over 260 square miles (12.35 of which is urban area with 12 districts and 85 neighborhoods), so how does one choose the best location to open their tech startup in such a big city?

1.2 The problem

Data that may contribute to determining the best area to open a tech startup might include: the proximity to relevant entities, such as training centers, universities and incubators the population density of the neighborhood; the distance to amenities, like restaurants, bars and cafés; the rental price; and proximity to transportation hubs, e.g. bus stops and taxis. This project aims to determine (and maybe rank) the best locations to open a tech startup in the city of Florianopolis based on these data.

1.3 Interest

This report can be of special interest for entrepreneurs interested in launching their startup, or considering the relocation of a pre-existing company, students who are considering their career choices, the teachers and professors who such students, real estate businesses, universities and the city council.

2. Data acquisition and cleaning

2.1 Data Sources

Data regarding the city's demographic information and public services offered by the city and the state was obtained through the governments' official channels and wikipedia; data about the city's businesses and the universities (including their locations) was obtained by queries through the Foursquare API.

Please note that the references used by this study will be included at the end of this document.

2.2 Data cleaning

Data for the universities' positions was obtained through the Foursquare API and put into a table, contained in a pandas DataFrame.

The data on the city's characteristics - names of the neighborhoods, the respective populations and the population ranking - was obtained through the city'sofficial Wikipedia page and put in a separate table, contained in a pandas DataFrame designed to house data about the different neighborhoods.

The latitude and longitude for each of the 85 neighborhoods in the city of Florianópolis were obtained through the API Open Cage Geocode, and added to the neighborhood's DataFrame.

The area for each of the neighborhoods was obtained through individual search in the official city hall’s website and stored in a .csv file. The data was then added to the neighborhoods’ table and later used for the calculation of the population density. The algorithm used to accomplish this bit was also included in the GitHub repository for this project.

With respect to the data about the venues’ locations and quantities, they were also obtained through the Foursquare API, and added to the neighborhoods’ table. Because the data was relatively low in volume and was readily available in trusted websites, there was no missing data in the database.

This is the city that is the object of this study, with its neighborhoods marked in blue:

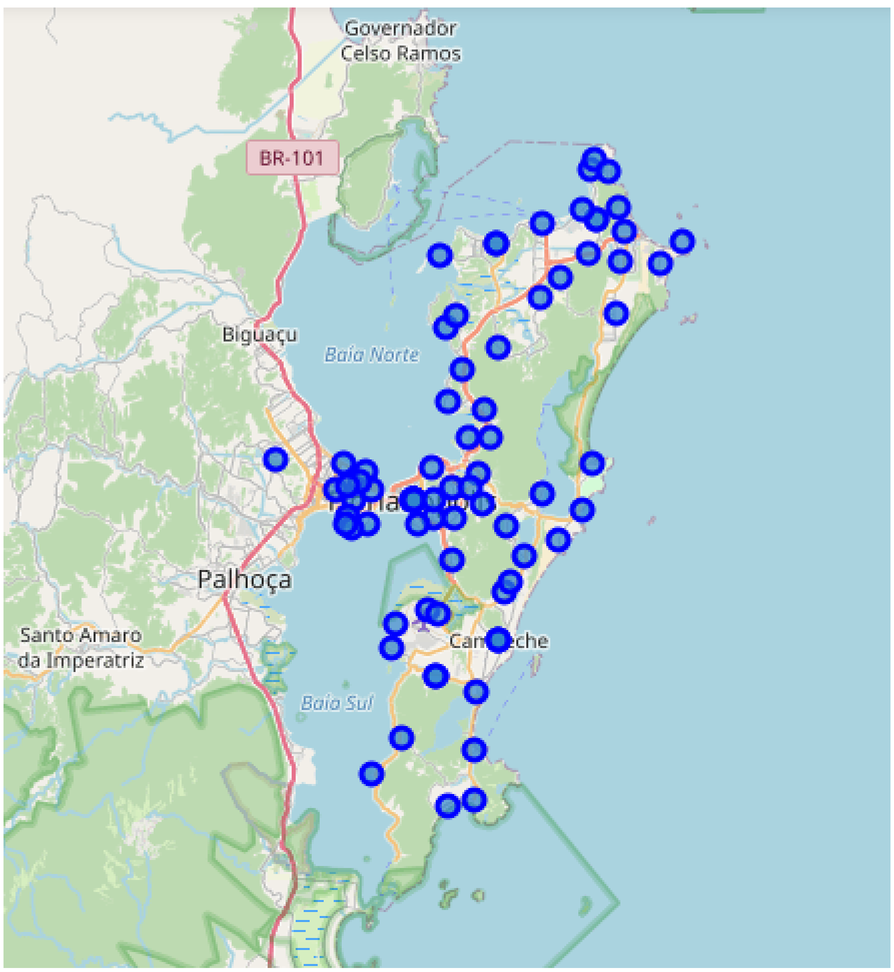

Figure 1. Florianópolis and its neighborhoods.

As you can see, the capital of the state of Santa Catarina is an island, with some of its territory in the continent. The metropolitan region was not considered in this study.

2.3 Feature selection

After data cleaning, there were 85 neighborhoods with 9 features – population density, number of bars in the neighborhood, number of restaurants, number of gyms, number of markets, number of bus stops, and the distances between the neighborhood and each of the 3 top universities in the city.

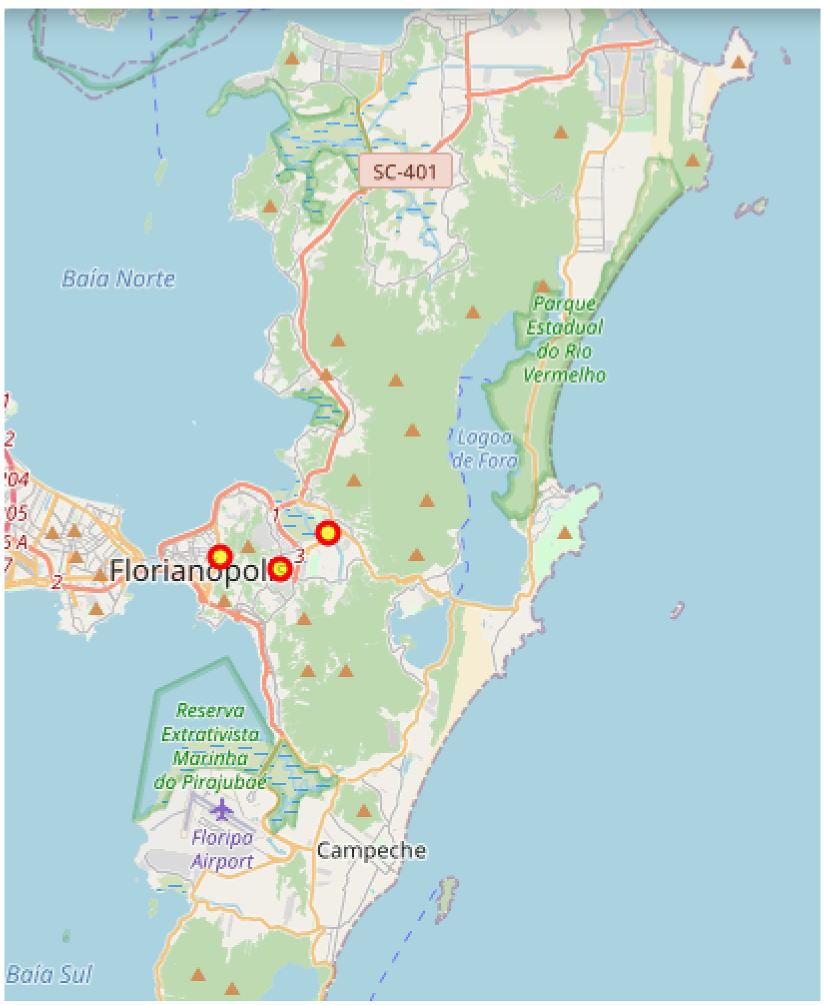

Figure 2. The location of the top 3 universities in the city

3. Methodology

Upon analysis of the population density, it was discovered that the density of places reported as the best places to live and/or work in the city (by locals) were far from the ideal – much higher than 100 inhabitants per km². Because of this, this feature was excluded from the study.

Some of the features selected for this analysis may seem unusual; one may ask themselves: but why is the number of bars and restaurants important in the location of a tech startup? As it happens, quality of life has been cited in a number of articles – the most proeminent of those were cited in the bibliography – as a key factor when choosing the address for a startup, and easy access to bars, markets and restaurants may be considered important in the daily life of startup workers. The same can be said of the access to transportation (buses being the major public transportation system in the city, this access can be measured by the number of bus stations nearby) and of the access to fitness centers (covered by the amount of gyms nearby).

After filtering for these results, the first rows of the dataframe for the neighborhoods looked like this:

	neighborhood	distance ufsc	distance ifsc	distance udesc	bars	restaurants	gyms	markets	bus stations
0	Centro	2.709233	0.819431	4.553951	50	50	47	25	9
1	Capoeiras	6.704180	4.720727	8.470627	50	21	26	6	9
2	Trindade	0.868572	2.015336	1.751967	50	34	25	8	5
3	Agronômica	2.512783	1.824074	3.151455	24	14	13	6	6
4	Saco dos Limões	1.732293	1.824608	3.801057	18	8	4	6	1

Figure 3. Neighborhoods' DataFrame with features.

In order to group the neighborhoods with similar characteristics a k-means clustering algorithm was employed in this dataset, as displayed below:

```python
# set number of clusters
kclusters = 4

floripa_grouped_clustering = df_pop.drop(['neighborhood', 'latitude', 'longitude'], 1)

# run k-means clustering
kmeans = KMeans(n_clusters = kclusters, random_state=0).fit(floripa_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

```
array([1, 3, 3, 2, 2, 2, 3, 3, 3, 2], dtype=int32)
```

```python
df_pop['cluster'] = kmeans.labels_
```

Figure 4. Clustering algorithm used in the analysis

This algorithm was used with different numbers of clusters (2 to 10), in order to determine the one that could group the neighborhoods with similar characteristics more accurately. Usually, descriptive statistics tools would be required to determine the uniformity of said clusters, but because we only have 85 rows in the database, an individual analysis of each cluster was considered enough – and more efficient – for the purposes of this study.

The most appropriate number of clusters – the one that best categorized the data – was 4.

4. Results

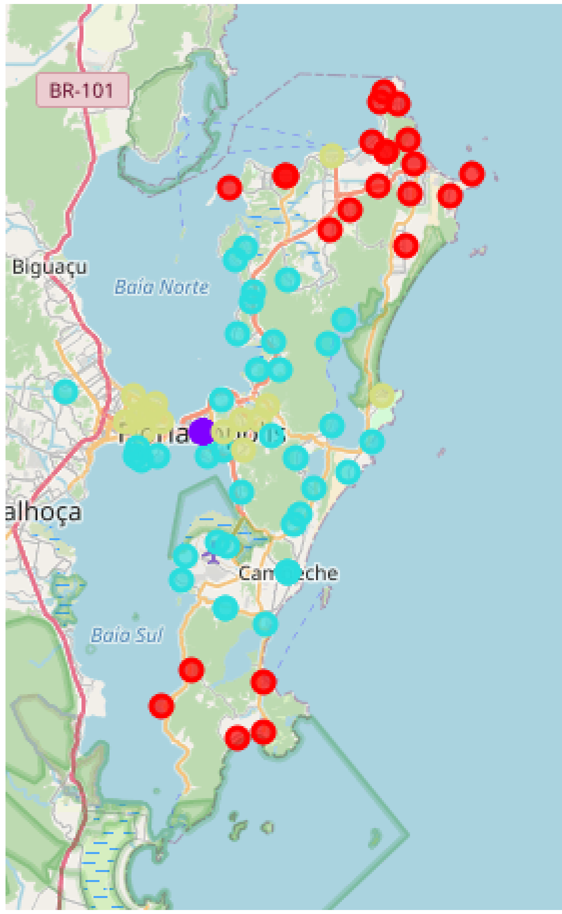

Figure 5. Neighborhoods after the clustering

The first cluster yielded by the algorithm has this appearance:

	neighborhood	distance ufsc	distance ifsc	distance udesc	bar	restaurante	academia	mercado	ponto de ônibus	cluster
10	Capivari	19.679754	20.585249	17.716058	6	8	5	7	1	0
15	São João do Rio Vermelho	16.788036	17.888995	14.761430	3	0	4	4	2	0
28	Ingleses Centro	21.582851	22.392932	19.655183	24	25	12	7	0	0
33	Santinho	21.288775	22.376551	19.259316	8	11	2	6	1	0
34	Ponta das Canas	24.200936	24.674262	22.440236	11	21	1	1	0	0
35	Vargem do Bom Jesus	19.060123	19.758953	17.193288	2	2	1	3	0	0
36	Armação	16.949528	17.637619	18.128988	9	5	2	1	0	0
37	Cachoeira do Bom Jesus Leste	21.316422	21.949212	19.476031	6	7	4	4	0	0

Figure 6. First cluster

As can be noticed on Figure 6, the neighborhoods present in the first cluster have in common large distances to the universities, and a low number of bus stations and other amenities. Naturally, these can not be considered suitable for a startup, according to the criteria established in this study.

The second cluster looks like this:

	neighborhood	distance ufsc	distance ifsc	distance udesc	bar	restaurante	academia	mercado	ponto de ônibus	cluster
0	Centro	2.709233	0.819431	4.553951	50	50	47	25	9	1
11	Tapera da Base	2.679143	0.763719	4.508555	50	50	50	27	9	1
41	Campeche Norte	2.679143	0.763719	4.508555	50	50	50	27	9	1
48	Ribeirão da Ilha[1]	2.679143	0.763719	4.508555	50	50	50	27	9	1
56	Rio Tavares do Norte	2.679143	0.763719	4.508555	50	50	50	27	9	1
60	Retiro	2.721909	0.835825	4.569069	50	50	47	26	9	1
70	Moenda	2.679143	0.763719	4.508555	50	50	50	27	9	1
76	Canto do Lamim	2.679143	0.763719	4.508555	50	50	50	27	9	1
77	Vargem de Fora	2.679143	0.763719	4.508555	50	50	50	27	9	1
79	Autódromo	2.679143	0.763719	4.508555	50	50	50	27	9	1

Figure 7. Second cluster

This cluster contains neighborhoods with a good amount of amenities, and relatively low distances to the universities. Let's see how it compares to the others.

	neighborhood	distance ufsc	distance ifsc	distance udesc	bar	restaurante	academia	mercado	ponto de ônibus	cluster
3	Agronômica	2.512783	1.824074	3.151455	24	14	13	6	6	2
4	Saco dos Limões	1.732293	1.824608	3.801057	18	8	4	6	1	2
5	Coqueiros	6.006306	4.279352	7.979264	21	20	16	2	4	2
9	Costeira do Pirajubaé	4.115277	4.920061	5.630187	5	4	3	1	0	2
13	Monte Verde	4.907345	5.918496	3.082996	7	5	3	0	1	2
17	Abraão	7.123377	5.246284	9.009980	26	8	14	2	5	2
19	Lagoa	8.271092	9.520870	6.229953	2	15	0	0	1	2
20	Saco Grande	6.408835	6.981250	4.833196	6	6	2	5	0	2
21	Córrego Grande	2.038580	4.078602	1.483266	24	7	7	2	2	2
26	Carianos	7.615009	7.857676	9.284115	8	6	4	3	3	2

Figure 8. Third cluster

In this cluster, one can notice that the neighborhoods have various disances to the universities, but a generally low number of amenities. These neigborhoods will also be excluded from the analysis.

	neighborhood	distance ufsc	distance ifsc	distance udesc	bar	restaurante	academia	mercado	ponto de ônibus	cluster
1	Capoeiras	6.704180	4.720727	8.470627	50	21	26	6	9	3
2	Trindade	0.868572	2.015336	1.751967	50	34	25	8	5	3
6	Monte Cristo	7.884722	5.856112	9.537097	50	33	24	8	7	3
7	Jardim Atlântico	7.758980	5.730976	9.123975	44	19	16	7	2	3
8	Itacorubi	2.473906	3.986870	0.513700	30	25	31	10	10	3
12	Estreito	5.535307	3.499551	7.184731	50	36	22	7	7	3
14	Balneário	6.149678	4.126147	7.538082	35	27	17	5	6	3
16	Canto	6.435640	4.390978	8.004857	50	39	25	9	10	3
18	Santa Mônica	1.456151	3.239256	0.629822	46	27	22	5	7	3
22	Canasvieiras	19.614088	19.947491	17.959644	46	36	9	9	1	3

Figure 9. Fourth cluster

Finally, in this last cluster there are neighborhoods with a reasonable number of amenities, and varied distances from the universities.

5. Difficulties

During the execution of this project, some difficulties arised, and they will be exposed below:

Wrong coordinates for one entry

For entry index 49, the coordinates for a neighborhood were 30km off. This was not discovered until the clustering step, during which one single cluster was generated with this neighborhood (Santo Antônio), and its entry accused no amenities nearby. The cause of the problem was pinpointed to the API query for the address. The problem was solved by inputing the correct entries manually, but an alternative solution could be querying the neighborhood by its full name (Santo Antônio de Lisboa).

Lack of a better API to measure distances

The API user for the positioning of the neighborhood was the OpenCage Geocoder, which, in addition to being free to use, has proven to be useful and accurate, but does not provide the user with road distances.

Area scraping

The data available for the areas of the neighborhoods was located in a state database. However, the system only accepts queries one-by-one, which means the DataFrame for the data had to be written entry by entry. The algorithm for the program used to build said DataFrame was included in the same GitHub depository as this report.

6. Discussion

As shown above, the best contenders for the title of best neighborhoods in Florianópolis to open a tech startup are the neighborhoods in clusters 2 and 4.

For the author, despite not being a zombie, it is really important to get as close as possible to those brains, which means 4.5 km may be too long of a distance between the company and one of the best universities in the country. In this case, the cluster that best fit the criteria is number 4; if we consider only the entries with distances of less than 4.5 km to any of the universities, we are left with the following:

neighborhood	latitude	longitude	distance ufsc	distance ifsc	distance udesc	bars	restaurants	gyms	markets	bus stops	cluster
Trindade	-27.589383	-48.522400	0.868572	2.015336	1.751967	50	34	25	8	5	3
Itacorubi	-27.581510	-48.504193	2.473906	3.986870	0.513700	30	25	31	10	10	3
Santa Mônica	-27.589928	-48.509703	1.456151	3.239256	0.629822	46	27	22	5	7	3
Pantanal	-27.608578	-48.521142	1.270104	2.680122	2.965065	30	20	5	5	6	3
Caieira	-27.596691	-48.535275	1.309972	0.781621	3.207379	50	50	15	14	0	3

Figure 10. Results filtered by distance to the universities

For the neighborhood ‘Caieira’, we have a number of bus stops equal to 0. That does not mean the neighborhood does not have any bus stops, but can mean that the neighborhood does not have bus stops in a radius of 1 km (or 0.62 miles) from the neighborhood center. In any case, let’s exclude it from the results.

In conclusion, the best neighborhoords to open a tech startup in Florianópolis, according to the criteria and analysis described above, are:

- Trindade
- Itacorubi
- Santa Mônica
- Pantanal

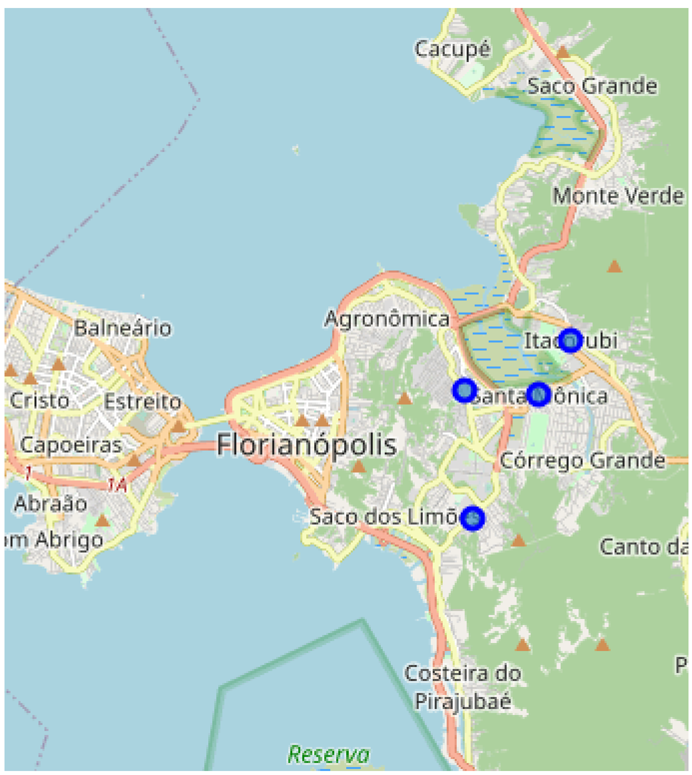

Figure 11. The best neighborhoods to open a tech startup

7. Possible improvements

The author considers the following points of this project as potentially important, as far as possible improvements are concerned:

- The distance measured using the gps coordinates of the neighborhoods does not, necessarily, reflect reality, because it is the linear distance, and not the road distance, which would required a paid API (or 85 manual researchs to Google Maps) to be obtained. In conclusion, this project could be improved by inputing the road distances instead of the linear distances.

- In the analysis of the venues described as ‘amenities’, all the venues that matched the description were considered. It is believed that the results could be more reliable by counting only the venues with good reviews.

- A different set of amenities could have been considered, affecting the results even further.

-Descriptive statistics can be used to analyze the clusters.

Bibliography

en.wikipedia.org. 2020. Florianópolis. [online] Available at: <https://en.wikipedia.org/wiki/Florian%C3%B3polis> [Accessed 1 May 2020].

pt.wikipedia.org. 2020. Florianópolis. [online] Available at: <https://pt.wikipedia.org/wiki/Florian%C3%B3polis> [Accessed 1 May 2020].

Ibge.gov.br. 2020. Florianópolis (SC) | Cidades E Estados | IBGE. [online] Available at: <https://www.ibge.gov.br/cidades-e-estados/sc/florianopolis.html> [Accessed 1 May 2020].

Sc.gov.br. 2020. Início - Governo Do Estado De Santa Catarina. [online] Available at: <https://www.sc.gov.br/> [Accessed 1 May 2020].

Pmf.sc.gov.br. 2020. Prefeitura De Florianópolis. [online] Available at: <http://www.pmf.sc.gov.br/> [Accessed 1 May 2020].