Capstone Project – The Battle of Neighborhoods

Clustering Florianópolis beaches

Natalia Assad

Introduction

The Florianópolis city is marked by its large number of beaches. These beaches vary in size and infrastructure. Some are considered deserted, with few shops around, where the option is just to enjoy nature. Others are already marked by having great infrastructure, have a variety of hotels, restaurants and clubs. This project aims to define the level of infrastructure for each beach in Florianópolis. For this, a clustering of the beaches will be performed based on the number of places found around each beach.

Data

For the selection of the beaches in Florianópolis, the name of the beaches provided by the Guia Floripa website was used (https://guiafloripa.com.br/turismo/praias).

It was also necessary to find the location data for each beach (latitude and longitude), for this reason the geopy tool was used.

Finally, we used data from the foursquare API, which returns a list of locations close to the location corresponding to the search, in this case, the beaches.

Methodology

The construction of the notebook took place in several stages. First, weeb scraping on the Guia Floripa website was necessary to obtain a list containing the name of the beaches in Florianópolis. After that, the geopy tool was used, in order to find the latitude and longitude of each beach contained in the list. The next step was to carry out a data treatment / cleaning, as the location tool searched for the wrong location of some beaches, returning beaches from outside the city of Florianópolis, which is not the scope of this project. Therefore, a manual correction of these points was necessary.

For a visualization of the beaches in Florianópolis, the folium library was used, which returned a map of the city of Florianópolis, with points that refer to the beaches that will

be analyzed. In this way, we were able to visually check the location of each beach to be studied.

Having correctly selected the beaches to be analyzed, the foursquare API was explored, where the search for places was carried out, given the latitude and longitude information of each beach contained in the dataframe. Through this API, it was possible to find the number of places around each beach, which will be used as a criterion to measure the infrastructure level of each beach.

Finally, the clustering of these beaches was carried out. For this, the K-means algorithm was used, which provides a classification of information according to the data itself, grouping information with similarities.

After that, the clusters found were examined, and an analysis of the information contained was developed.

The following is a description of each step.

- **Weeb Scrapping**

For the realization of weeb scrapping it was necessary to import the Beautiful Soup Python library, this library is used for the analysis of HTML documents, through which it is possible to extract HTML data.

Through weeb scraping, it was possible to list the names of Florianópolis beaches described on the Guia Floripa website.

- **Geopy**

Geopy is a tool capable of converting physical addresses to geographic locations, such as latitude and longitude. In this work, the name of the beach and the state to which it belongs is provided as input data, so the tool will return the latitude and longitude of this beach.

With this tool, we were able to build a dataframe that contains the name of the beach with its respective latitudes and longitudes.

FIGURe 1 -Dataframe

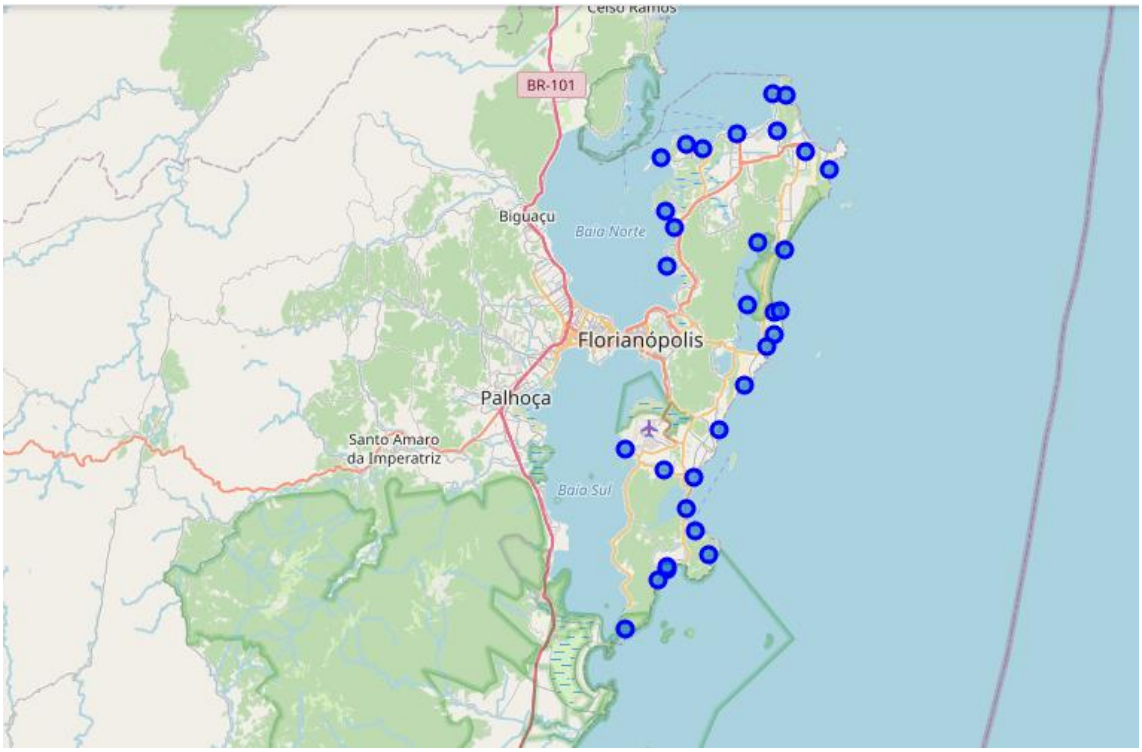
	Praias	Latitude	Longitude
0	Cachoeira do Bom Jesus, SC	-27.426531	-48.423675
1	Cacupé, SC	-27.537201	-48.524557
2	Canasvieiras, SC	-27.429567	-48.460244
3	Daniela, SC	-27.448954	-48.531250
4	Ingleses, SC	-27.443666	-48.397142
5	Jurerê Internacional, SC	-27.438203	-48.507074
6	Jurerê Tradicional, SC	-27.441564	-48.491754
7	Lagoinha, SC	-27.252150	-50.557237
8	Ponta das Canas, SC	-27.396429	-48.427428
9	Praia Brava, SC	-27.397613	-48.415825
10	Praia do Forte, SC	-26.168858	-48.534342
11	Praia do Santinho, SC	-27.458361	-48.375006
12	Sambaqui, SC	-27.492403	-48.526831
13	Santo Antônio de Lisboa, SC	-27.506151	-48.518870

This dataframe will be used both to plot the location of the beaches on the map generated by the Folium, and to find the nearby places for each location in the foursquare API.

- **Folium**

Folium Map is a Python library used to visualize geospatial data, you only need the location parameter to create your map. This parameter receives a latitude-longitude pair. When creating the map, it is still possible to add points that you want to make an analysis. In this project, the map created was of the city of Florianópolis, and the added points refer to each beach in that region. When calling the map, we have the following visualization.

FIGURE 2 – Florianópolis Map



- **Foursquare API**

With the proper location of each beach, we will explore its surroundings. For that, we will need to access the foursquare API. You will need a developer registration on the site in order to obtain the access credentials, they are Client_ID and Client_Secret.

With that, we can access the API and search for the necessary information. For our case, we will use the get –venues-explore method.

A function will be defined in order to access the foursquare API and find information about places close to each beach. The limit of 30 venues per beach was defined, within a radius of 500 meters. In such a way, that the return of the function will give us a dataframe, which contains the place found next to each beach, with the description of the latitude and longitude of each establishment, the name of the establishment, and the category of the establishment. The following figure shows what is returned by the function.

FIGURE 3 – DataFrame with venues

	Praias	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Cachoeira do Bom Jesus, SC	-27.426531	-48.423675	Confiance Restaurante	-27.426191	-48.424464	Buffet
1	Cachoeira do Bom Jesus, SC	-27.426531	-48.423675	Panificadora União	-27.426834	-48.423778	Bakery
2	Cachoeira do Bom Jesus, SC	-27.426531	-48.423675	Santa Pizzaria e Restaurante	-27.426198	-48.424431	Pizza Place
3	Cachoeira do Bom Jesus, SC	-27.426531	-48.423675	Mercado Petroski	-27.426307	-48.425374	Grocery Store
4	Cachoeira do Bom Jesus, SC	-27.426531	-48.423675	Lanches Do Jorge	-27.425907	-48.424526	Diner
...	...	...	...	...	...	...	...
298	Ribeirão da Ilha, SC	-27.703893	-48.528631	Boteco Chamarrita	-27.704251	-48.527399	Burger Joint
299	Tapera, SC	-27.686813	-48.564063	Praia da Tapera	-27.686585	-48.564656	Beach
300	Tapera, SC	-27.686813	-48.564063	Academia Equilíbrio	-27.687717	-48.565420	Gymnastics Gym
301	Tapera, SC	-27.686813	-48.564063	Padaria Casa dos Pães	-27.686263	-48.561671	Bakery
302	Tapera, SC	-27.686813	-48.564063	Tapera, Florianópolis	-27.686277	-48.561667	Beach

As we want to analyze only the number of establishments around each beach first, we will use the groupby method, to group each beach, and the count method, to return only the number of establishments. Obtaining, the following dataframe.

FIGURE 4 – Dataframe for clustering

Praias	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Barra da Lagoa, SC	30	30	30	30	30	30
Cachoeira do Bom Jesus, SC	5	5	5	5	5	5
Cacupé, SC	8	8	8	8	8	8
Canasvieiras, SC	30	30	30	30	30	30
Daniela, SC	6	6	6	6	6	6
Ingleses, SC	7	7	7	7	7	7
Jurerê Internacional, SC	12	12	12	12	12	12
Jurerê Tradicional, SC	30	30	30	30	30	30
Ponta das Canas, SC	11	11	11	11	11	11
Praia Brava, SC	13	13	13	13	13	13
Praia Mole, SC	12	12	12	12	12	12
Praia da Armação, SC	7	7	7	7	7	7

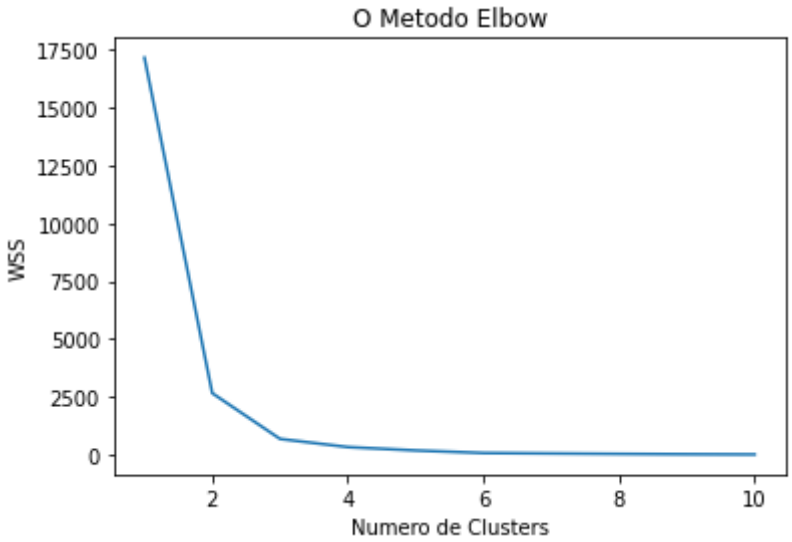
- **Clustering**

The algorithm used in the classification will be K-means, for that it will be necessary to import it from the Sklearn library.

Such an algorithm provides a classification based on analysis and comparisons between numerical values. In this way, the algorithm will automatically provide an automatic classification without the need for any human supervision, that is, without any existing pre-classification. Because of this feature, K-Means is considered to be an unsupervised data mining algorithm.

To perform the clustrerization by the K-means method it is necessary to inform the number of clusters that is desired. To find the number of clusters that best fits the data, the Elbow method was used.

FIGURE 5 –Elbow Method



As you can see from the image above, the number of clusters that best fits our data set is 3. Therefore, the number of 3 clusters to perform the clustering was defined.

The figure below shows the algorithm used to perform the clustering.

FIGURE 6 – Algorithm K-means

```
[ ]  # set number of clusters
     kclusters = 3

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

     # check cluster labels generated for each row in the dataframe
     kmeans_labels = kmeans.labels_
```

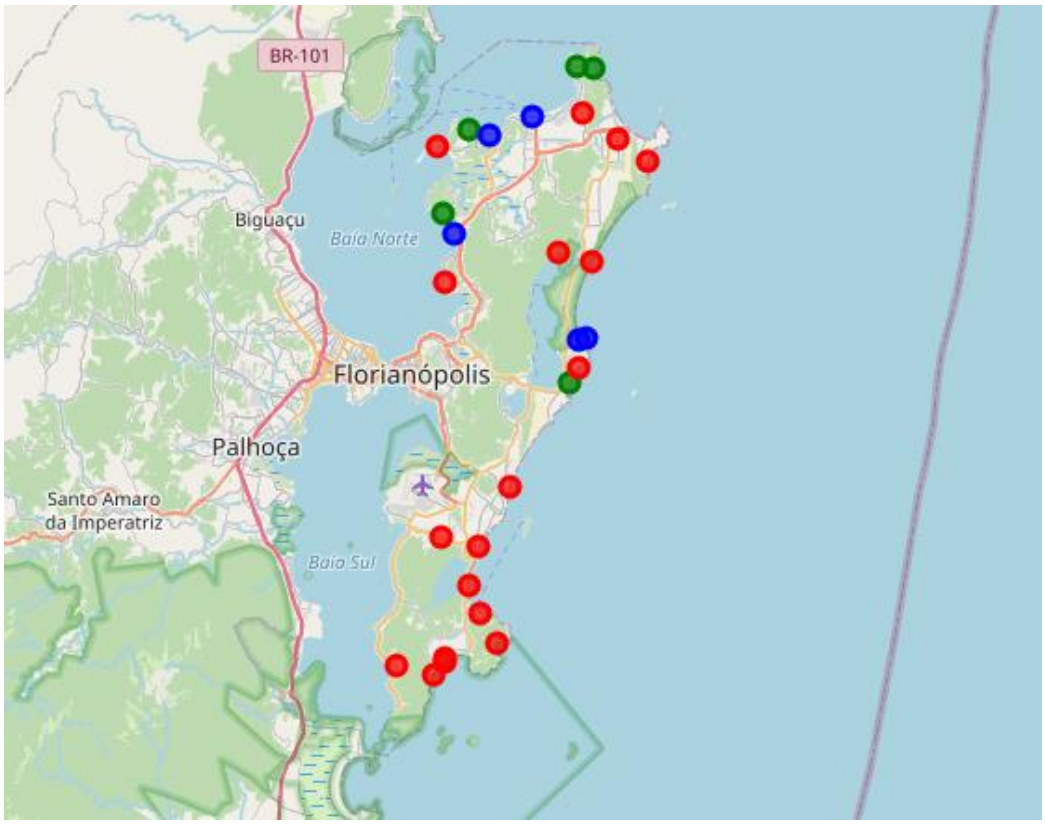
The dataframe below shows the result of the clustering, where for each beach it is possible to check which cluster it belongs to.

FIGURE 7 – Dataframe clustered

	Praias	Cluster Label	Venue Category	Latitude	Longitude
0	Barra da Lagoa, SC	1	30	-27.574778	-48.425835
1	Cachoeira do Bom Jesus, SC	0	5	-27.426531	-48.423675
2	Cacupé, SC	0	8	-27.537201	-48.524557
3	Canasvieiras, SC	1	30	-27.429567	-48.460244
4	Daniela, SC	0	6	-27.448954	-48.531250
5	Ingleses, SC	0	7	-27.443666	-48.397142
6	Jurerê Internacional, SC	2	12	-27.438203	-48.507074
7	Jurerê Tradicional, SC	1	30	-27.441564	-48.491754
8	Ponta das Canas, SC	2	11	-27.396429	-48.427428
9	Praia Brava, SC	2	13	-27.397613	-48.415825
10	Praia Mole, SC	2	12	-27.603133	-48.433334
11	Praia da Armação, SC	0	7	-27.736035	-48.507903
12	Praia da Galheta, SC	0	2	-27.593527	-48.426915

Finally, the map of Florianópolis is shown again, with the respective beaches. Only now, it will be possible to identify which grouping each beach belongs to. The beaches will be identified by colored dots, which identify which cluster the beach belongs to.

FIGURA 8 – Florianopolis Map Clustered



The red dots represent the beaches that are part of cluster 0, the blue dots represent the beaches that are part of cluster 1, and the green dots represent the beaches that belong to cluster 2.

Results

By examining each cluster individually, we can identify the clustering pattern of the clusters.

For cluster 0, it is noted that the beaches with the fewest establishments found in the surroundings are concentrated. The number of establishments varies from 1 to 8.

FIGURA 9 – Cluster 0

	Praias	Cluster Label	Venue Category	Latitude	Longitude
1	Cachoeira do Bom Jesus, SC	0	5	-27.426531	-48.423675
2	Cacupé, SC	0	8	-27.537201	-48.524557
4	Daniela, SC	0	6	-27.448954	-48.531250
5	Ingleses, SC	0	7	-27.443666	-48.397142
11	Praia da Armação, SC	0	7	-27.736035	-48.507903
12	Praia da Galheta, SC	0	2	-27.593527	-48.426915
13	Praia da Lagoinha do Leste, SC	0	1	-27.774022	-48.486880
14	Praia da Solidão, SC	0	3	-27.793965	-48.533928
16	Praia do Campeche, SC	0	4	-27.671869	-48.477448
17	Praia do Forte, SC	0	4	-26.168858	-48.534342
18	Praia do Gravatá, SC	0	4	-26.841915	-48.631192
19	Praia do Matadeiro, SC	0	6	-27.754367	-48.498950
20	Praia do Morro das Pedras, SC	0	4	-27.710701	-48.500155
21	Praia do Moçambique, SC	0	2	-27.524143	-48.417212

These beaches are considered more desert, with a smaller number of establishments for you to get to know and enjoy. Here the tourist who seeks tranquility, and just wants to enjoy the sea and nature, will find a good option to stay.

Cluster 1, on the other hand, concentrates the beaches with the greatest choice of establishments. In this cluster, the beaches with the largest number of establishments around are grouped together.

FIGURE 10 – Cluster 1

	Praias	Cluster Label	Venue Category	Latitude	Longitude
0	Barra da Lagoa, SC	1	30	-27.574778	-48.425835
3	Canasvieiras, SC	1	30	-27.429567	-48.460244
7	Jurerê Tradicional, SC	1	30	-27.441564	-48.491754
26	Prainha da Barra da Lagoa, SC	1	30	-27.574137	-48.421103
29	Santo Antônio de Lisboa, SC	1	30	-27.506151	-48.518870

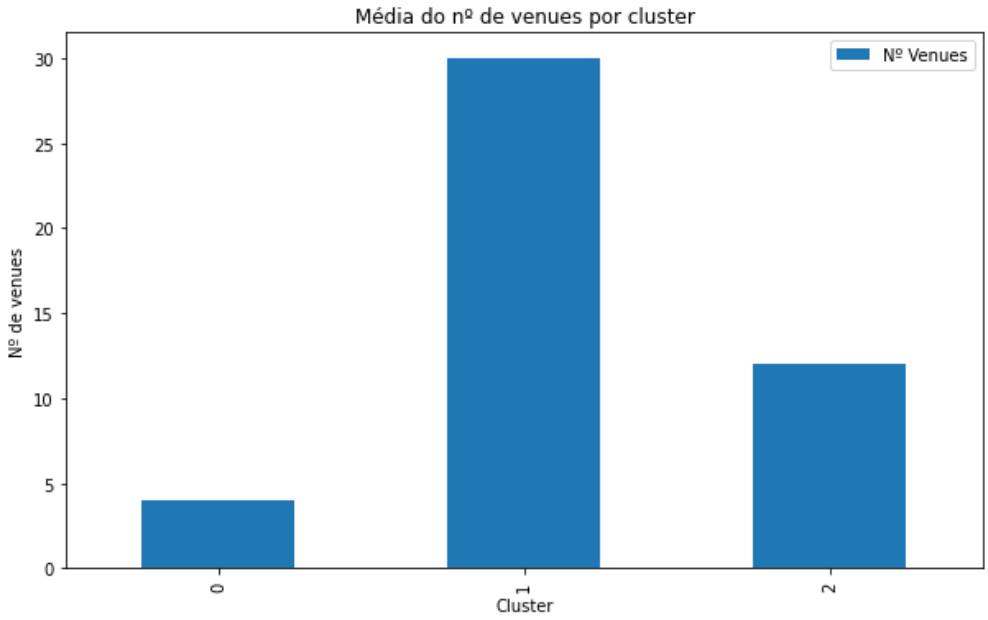
These beaches present a good option for the tourist who wants to combine the privilege of a beach to enjoy nature, but who also does not give up the activities that a city can offer, from restaurants to clubs.

Cluster 2 brought together the intermediate beaches, which are between the desert beaches and the beaches with the greatest infrastructure. The number of establishments varies from 10 to 17.

FIGURA 11 – Cluster 2

	Praias	Cluster Label	Venue Category	Latitude	Longitude
6	Jurerê Internacional, SC	2	12	-27.438203	-48.507074
8	Ponta das Canas, SC	2	11	-27.396429	-48.427428
9	Praia Brava, SC	2	13	-27.397613	-48.415825
10	Praia Mole, SC	2	12	-27.603133	-48.433334
15	Praia de Naufragados, SC	2	10	39.873143	-8.972437
28	Sambaqui, SC	2	17	-27.492403	-48.526831

The histogram below shows the average number of establishments per cluster.



Conclusion

From this analysis, it was found that in the city of Florianópolis it is possible to find 3 types of beaches based on infrastructure. More desert beaches, with few establishments around. Grouping 19 beaches with these characteristics.

More structured beaches, where there is a greater number of establishments to visit around the beach. Here the beaches of Barra da Lagoa, Canasvieiras, Jurerê Tradicional, Santo Antônio de Lisboa and Prainha da Barra da Lagoa stand out.

And finally, the beaches considered to be of medium infrastructure, in which they present a certain number of establishments, but which do not reach the number to be considered more structured beaches.