FINAL PROJECT –SOUVENIR STORE IN COPACABANA

INTRODUCTION

Copacabana with its world-famous beach attracts a large amount of tourists all year round. In addition, there are different events that contribute to further increase the capacity of hotels at certain times of the year.

So, combining tourist relevance with all the bohemian air that surrounds this neighborhood, a group of artistic artists expressed an interest in opening a business in Copacabana. As this we have the question to be solved: What is the best place to open a souvenir shop in Copacabana?

During the business understanding stage, the specifics of the business and the desire of new entrepreneurs became clear.

The new establishment will sell products made by the store owners. These pieces have a differentiated design which ends up leading to an increase in the price of the products. Thus, it is not necessary to take into account the number of stores that already exist in the same industry as they will not be competitors.

We noticed a large concentration of tourists in the vicinity of the hotels, so we opted for machine learning k-means to group the hotels and thus, check the surroundings where the hotels are more concentrated.

Hotel data will be collected through Foursquare which is a web service based on geo-location. This service provides various information about the location searched. After collecting, cleaning and filtering the data, we noticed that there is a rating for each hotel. This ranking of hotels should attract a greater number of tourists to the hotel that is best rated.

The objective of the project is to define the surroundings where there is a greater concentration of hotels, in addition to this, we will evaluate the averages of the marks awarded in each group of hotels.

MATERIALS AND METHODS

Data were collected through the Foursquare web service, which is based on geo-location. It is possible to access place data through the available APIs.

The project needs to define an area of reach from a central location in the Copacabana neighborhood. The address chosen was R Domingos Ferreira 6 and the range was 2,000 meters.

As Foursquare needs latitude and longitude coordinates, the Nominatim geolocalizer was used for this purpose.

The Foursquare API returns the data in JSON format.


```
{'meta': {'code': 200, 'requestId': '6085570b93d76957ffab6b92'},
 'notifications': [{'item': {'unreadCount': 0}, 'type': 'notificationTray'}],
 'response': {'venues': [{'categories': [{'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/travel/hotel_',
       'suffix': '.png'},
      'id': '4bf58dd8d48988d1fa931735',
      'name': 'Hotel',
      'pluralName': 'Hotels',
      'primary': True,
      'shortName': 'Hotel'}],
    'hasPerk': False,
    'id': '4b058720f964a520128122e3',
    'location': {'address': 'Av. Atlântica, 2600',
     'cc': 'BR',
     'city': 'Rio de Janeiro',
     'country': 'Brasil',
     'distance': 76,
     'formattedAddress': ['Av. Atlântica, 2600',
      'Rio de Janeiro, RJ',
      '22041-001'],
     'labeledLatLngs': [{'label': 'display',
       'lat': -22.9722307,
       'lng': -43.1858418}],
     'lat': -22.9722307,
     'lng': -43.1858418,
     'postalCode': '22041-001',
     'state': 'RJ'},
    'name': 'JW Marriott Hotel Rio de Janeiro',
    'referralId': 'v-1619351307',
    'venuePage': {'id': '182987381'}},
```


In the pre-processing stage, the data was transformed into a 50-row, 19-column dataframe format. After analyzing the information, columns with unnecessary information for the project were deleted. Below the dataframe with the information pertinent to the hotels:

	id	name	categories	address	hotelLatitude	hotelLongitude
0	4b058720f964a520128122e3	JW Marriott Hotel Rio de Janeiro	[{'id': '4bf58dd8d48988d1fa931735', 'name': 'H...	Av. Atlântica, 2600	-22.972231	-43.185842
1	4bf3fa3dcad2c928b0589b99	Grande Hotel Canadá	[{'id': '4bf58dd8d48988d1fa931735', 'name': 'H...	Av. N.Sa. de Copacabana, 687	-22.971394	-43.187073
2	4f084d40e4b0a2229aaa3317	Olinda Rio Hotel	[{'id': '4bf58dd8d48988d1fa931735', 'name': 'H...	Av. Atlântica, 2230	-22.970240	-43.182959
3	4b058720f964a520048122e3	Majestic Rio Palace Hotel	[{'id': '4bf58dd8d48988d1fa931735', 'name': 'H...	R. Cinco de Julho, 195	-22.971066	-43.190163
4	4b0fe5faf964a520026623e3	Hotel Astoria Palace	[{'id': '4bf58dd8d48988d1fa931735', 'name': 'H...	Av. Atlantica, 1866	-22.968334	-43.180009

By viewing the data on the map below, it was possible to check a hotel located in the Botafogo neighborhood. As the scope of the project involves only the Copacabana neighborhood, the data for this hotel has been deleted from the database.

Map of the hotels in Copacabana contained in the dataframe of this project

The project needs information regarding the marks awarded to each hotel. To obtain this information, a new search in the Foursquare API was necessary. At this time, a limitation in the use of the Foursquare service by a free registration user, made data collection impossible. As the database now has 49 hotels, 49 queries were required to search for the notes. After repeated consultations, it was no longer possible to carry out searches. As a solution, the result of the API was stored in a txt file.

It is worth mentioning that five hotel evaluation notes are missing from the API. In order not to discard data, the missing notes were included in the database after manual consultations with hotel review sites on the web.

The dataframe presents the following data stored after the previous steps of cleaning, filtering and formatting the data:

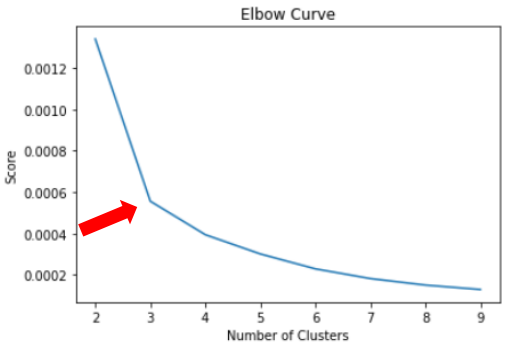
	id	name	categories	address	hotelLatitude	hotelLongitude	rating
0	4b058720f964a520128122e3	JW Marriott Hotel Rio de Janeiro	Hotel	Av. Atlântica, 2600	-22.972231	-43.185842	8.0
1	4bf3fa3dcad2c928b0589b99	Grande Hotel Canadá	Hotel	Av. N.Sa. de Copacabana, 687	-22.971394	-43.187073	6.3
2	4f084d40e4b0a2229aaa3317	Olinda Rio Hotel	Hotel	Av. Atlântica, 2230	-22.970240	-43.182959	6.6
3	4b058720f964a520048122e3	Majestic Rio Palace Hotel	Hotel	R. Cinco de Julho, 195	-22.971066	-43.190163	6.6
4	4b0fe5faf964a520026623e3	Hotel Astoria Palace	Hotel	Av. Atlantica, 1866	-22.968334	-43.180009	6.7

DATA ANALYSIS - DEVELOPMENT

The project seeks to solve the problem presented using machine learn K-means to group hotels in Copacabana into clusters.

The K value was defined using two metrics: Elbow and Silhouette.

The Elbow graph allows you to visualize the best K value as 3 because it is the value of the x-axis where the graph shows the “elbow”.



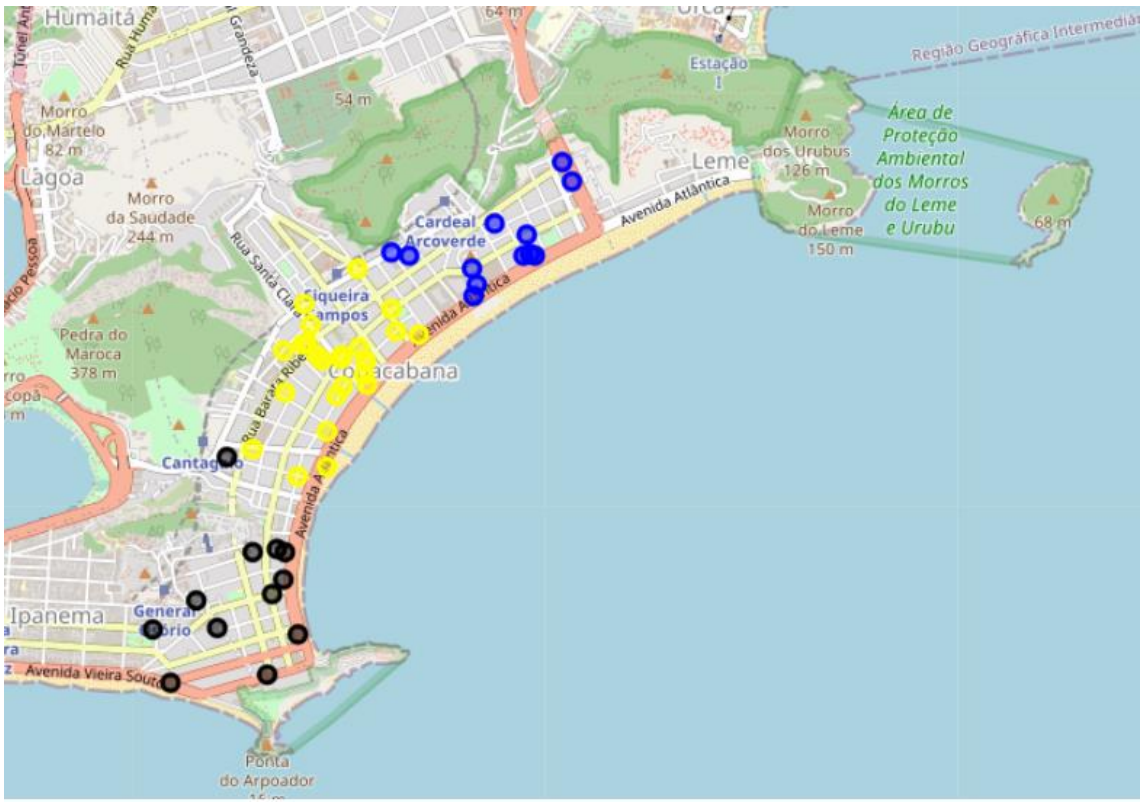
The highest value of the Silhouette metric indicates the best corresponding K value: K = 3

Número de Cluster	Silhouette
2	0,5490
3	0,5643
4	0,4730
5	0,4554
6	0,4532
7	0,4439
8	0,4494
9	0,4716

The table below shows the values for the cluster data:

	Contagem de hotéis	Média de notas dos hotéis
Cluster 0	12	6,43
Cluster 1	12	7,40
Cluster 2	25	6,93

Hotel clusters in Copacabana



CONCLUSION

The project aims to determine the best place for the opening of the souvenir shop. Through the k-means clustering method, the hotels were grouped in order to define a greater concentration of hotels and to evaluate the averages of the marks attributed to the hotels of each cluster.

Copacabana hotels were grouped into three clusters. Two of these clusters had a similar concentration of hotels. Another cluster, represented in yellow on the map below, represents more than twice the number of hotels compared to the other clusters. The best choice for the installation of the new trade was evidenced in this cluster (yellow color on the map).

Analyzing the average hotel scores by cluster, we noticed that cluster 1 (black on the map) presents a concentration of hotels that are better evaluated than the others. A greater difference is perceived if we compare the results of cluster 1 (black) and cluster 0 (blue).

The choice of the location of the new store is undoubtedly in cluster 2 due to its greater concentration of hotels in its surroundings.

The average of the grades ended up not being a relevant data for choosing the location of the store.

REFERENCES

https://learning.oreilly.com/learning-paths/learning-path-python/9781492025443/9781449369880-/ch01.html

https://developer.foursquare.com/docs/places-api/endpoints/