IBM Data Science Capstone Project

Finding potential locations for a touristic bar

Steve Loveday

October 2020

Lima, Perú.

1 Introduction

The project takes place in Lima. Lima is the capital and the largest city of Peru, where almost 10 million people live. Lima is located on the central coast of the country, overlooking the Pacific Ocean.
In 2018, of the total of foreign tourists who visited Peru, 72.4% visited Lima, the main places visited by foreigners being Miraflores: 69.1%, Downtown Lima: 62.7%, Barranco: 26.7% and San Isidro: 18.9%.

Business Problem.

The purpose of this project is to find the most appropriate location for a thematic bar mainly focused on tourists. It must be close to points of interest or hotels and away from other bars. Lima has 43 districts, two of them are the districts of interest selected, Miraflores and Barranco for being the most striking tourist districts for their cultural diversity, establishments, and entertainment as well as their magnificent view of the coast.
Considering these requirements, we can create a map to obtain the best solution.

2 Data

For our analysis, we needed to see the following aspects:

- A list of districts as well as statistical information on tourism and places of interest.
- Determine the location of bars and hotels to be analyzed.
- It is necessary to have the exact coordinates of the places defined by the client.

The data will be extract or generate from reliable sources.

- The list of districts and information of interest of the city of Lima will be obtained from Peruvian webs such a Prom Peru https://www.promperu.gob.pe/ and the Foreign Trade and Tourism Ministry https://www.gob.pe/mincetur

- The number and locations of the venues of interest will be obtained using Foursquare Developers Access: https://foursquare.com/
- The coordinates of the districts or neighborhoods defined by the client will be obtained from Geopy Geocoders: https://pypi.org/project/geopy/

3 Methodology

For this project, we will require the location of both districts, as well as the location of the hotels and bars in the area.

For the required data, the coordinates of the districts will be obtained from the Geopy Geocoders libraries.

	District	Latitude	Longitude
0	Barranco	-12.143959	-77.020268
1	Miraflores	-12.121498	-77.025906

To appreciate the obtained information, the results will be displayed using the Folium Package, which uses OpenStreetMap technology.

For the exploration of the locations of hotels and bars, we will use the foursquare API. The first venue of interest will be the locations of tourist hotels. A 1500 meters radius will be assigned.

The data obtained will be examined and relevant information transforms into a dataframe.

	name	categories	address	crossStreet	lat	lng	labeledLatLngs	distance	postalCode	cc
0	Hotel B	Hotel	Av. Sáenz Peña 204	San Martín	-12.142954	-77.023266	[{'label': 'display', 'lat': -12.1429543860206...	344	15063	PE
1	Hotel B	Hotel	Av. San Martin	NaN	-12.144337	-77.018891	[{'label': 'display', 'lat': -12.1443374363514...	155	NaN	PE
2	Hotel Nirvana	Hotel	Av. Paseo De La República 6315	NaN	-12.131778	-77.021658	[{'label': 'display', 'lat': -12.1317781225687...	1364	NaN	PE
3	JW Marriott Hotel Lima	Hotel	Malecon De La Reserva 615, Miraflores	NaN	-12.131734	-77.029395	[{'label': 'display', 'lat': -12.1317336, 'lng...	1684	15074	PE
4	JW Marriott - Hotel Bar	Hotel Bar	Ave. Larco	Mco de la Reserva	-12.131676	-77.029451	[{'label': 'display', 'lat': -12.1316761578210...	1693	Peru	PE
...	...	...	...	...	...	...	...	...	...	...
86	Hotel El Condado	Hotel	NaN	NaN	-12.123867	-77.027652	[{'label': 'display', 'lat': -12.1238674964387...	325	NaN	PE
87	Hotel Eaperanza	Hotel	Eaperanza 348	NaN	-12.119885	-77.028359	[{'label': 'display', 'lat': -12.119885, 'lng'...	321	NaN	PE
88	Hotel Shell	Hotel	NaN	NaN	-12.122921	-77.028739	[{'label': 'display', 'lat': -12.122921, 'lng'...	346	NaN	PE
89	Hotel Maria Luisa	Hotel	Pasaje Tello 241	NaN	-12.121810	-77.028323	[{'label': 'display', 'lat': -12.1218101043319...	265	NaN	PE
90	Hotel las palmas	None	NaN	NaN	-12.121220	-77.029417	[{'label': 'display', 'lat': -12.1212201564626...	383	NaN	PE

91 rows × 16 columns

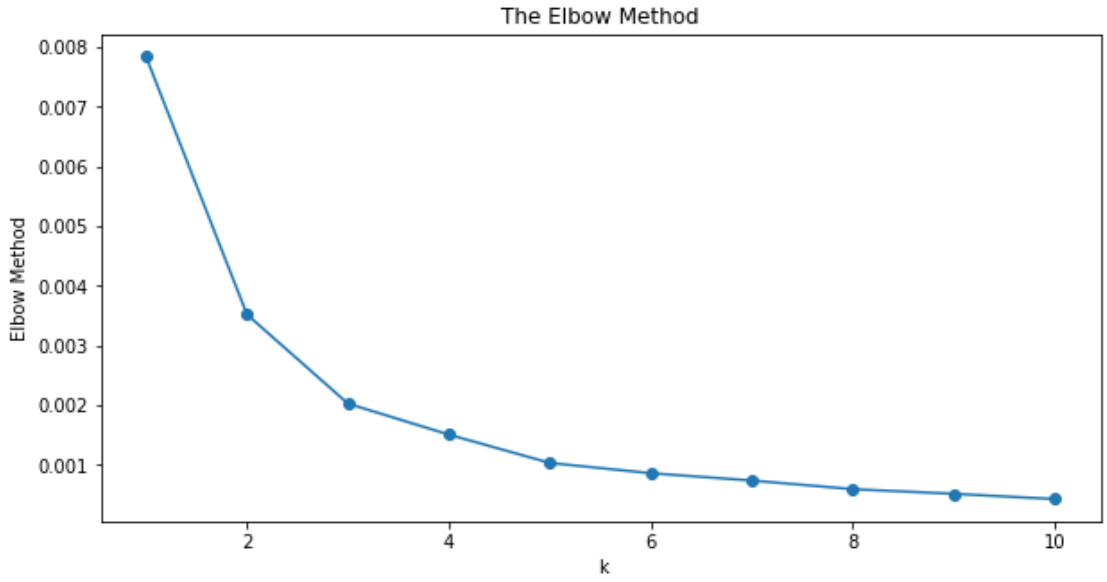
We can visualize the location of the hotels in Miraflores and Barranco districts.

We will focus on relevant information to get the dataset ready to be able to apply unsupervised machine learning technique K-means clustering for the creation of the hotel clusters.

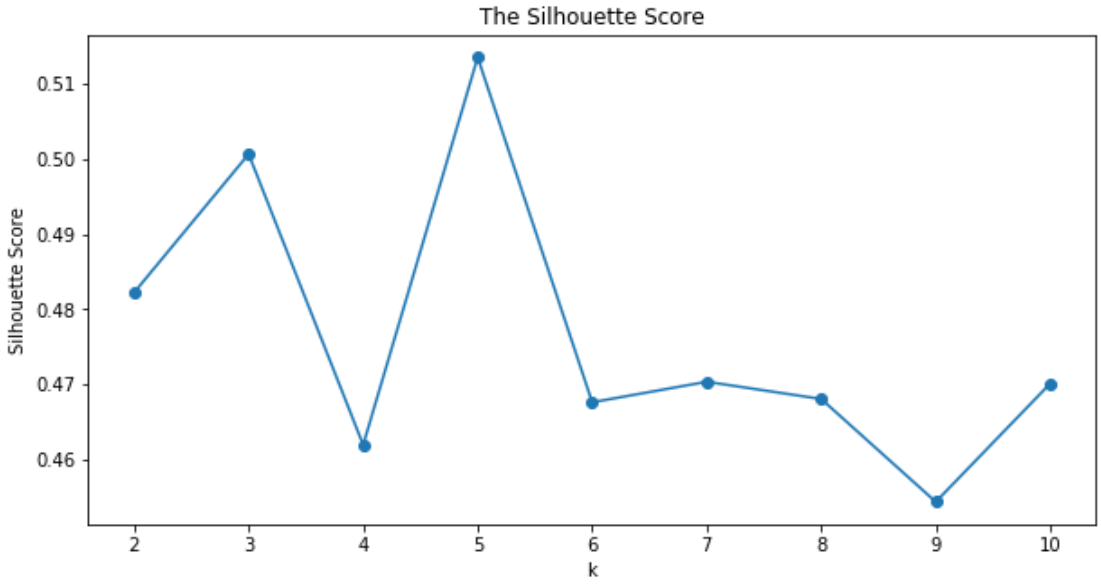
	Name	Latitude	Longitude
0	Hotel B	-12.142954	-77.023266
1	Hotel B	-12.144337	-77.018891
2	Hotel Nirvana	-12.131778	-77.021658
3	JW Marriott Hotel Lima	-12.131734	-77.029395
4	JW Marriott - Hotel Bar	-12.131676	-77.029451
...	...	...	...
86	Hotel El Condado	-12.123867	-77.027652
87	Hotel Eaperanza	-12.119885	-77.028359
88	Hotel Shell	-12.122921	-77.028739
89	Hotel Maria Luisa	-12.121810	-77.028323
90	Hotel las palmas	-12.121220	-77.029417

91 rows × 3 columns

For a better performing of the k-means clustering, we need to decide the number of clusters that are better to use. We will use the combined method of the Elbow Method and the Silhouette Score to choose the optimal number of clusters.



```
Max k: 0.5134341097255182
```



The Elbow Method is picking the "elbow" of the curve as the number of clusters to use, and the Silhouette Score is the highest score. As we can see from both graphs the optimal number of clusters is 5.

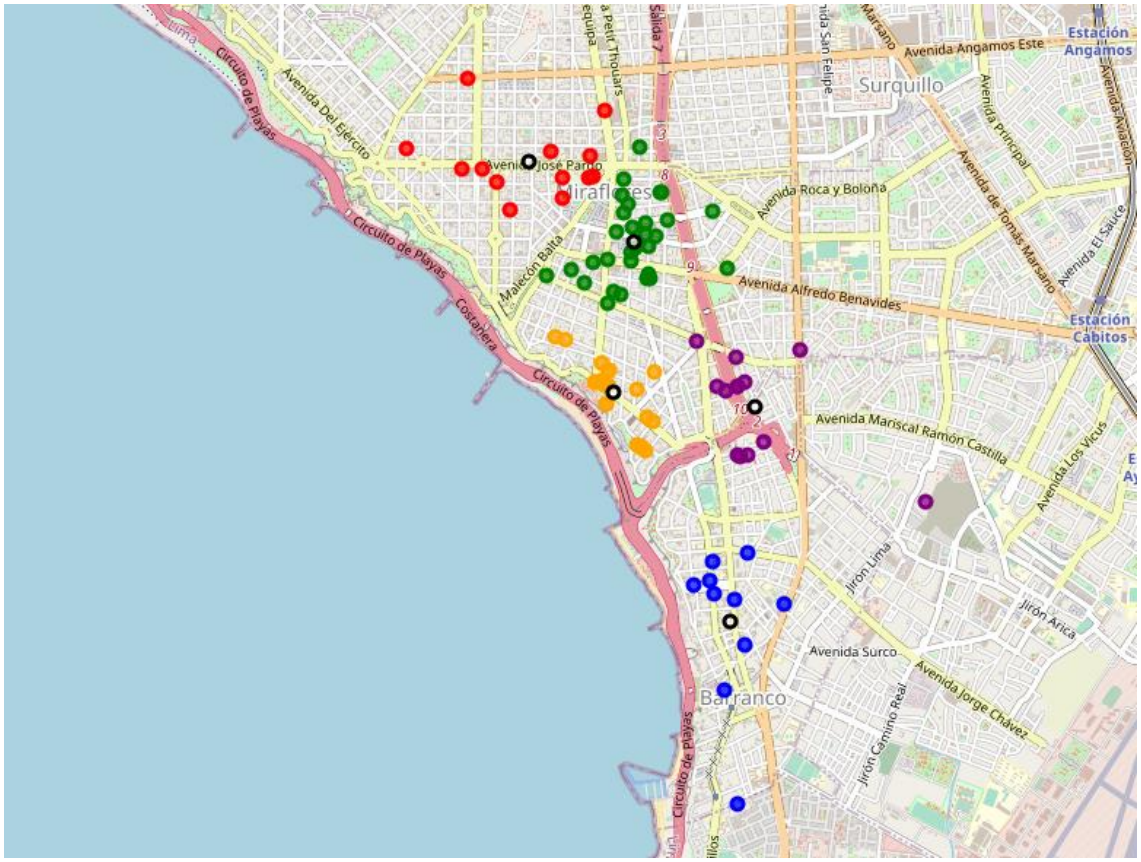
Now we can run k-means clustering. Once K-means algorithm is performed it assigns a label to each venue so we can add those to the dataset.

	Name	Latitude	Longitude	K_clusters
0	Hotel B	-12.142954	-77.023266	2
1	Hotel B	-12.144337	-77.018891	2
2	Hotel Nirvana	-12.131778	-77.021658	0
3	JW Marriott Hotel Lima	-12.131734	-77.029395	4
4	JW Marriott - Hotel Bar	-12.131676	-77.029451	4
5	Hotel apurimac	-12.141363	-77.021026	2
6	Bayview Hotel	-12.130876	-77.029193	4
7	Hotel Antu Suites	-12.146686	-77.021174	2
8	Bar Hotel B	-12.143727	-77.023001	2
9	Casa Falleri Boutique Hotel	-12.144083	-77.021830	2
10	Hotel	-12.135748	-77.021595	0
11	Hotel Arqueólogo Cusco	-12.143210	-77.024170	2

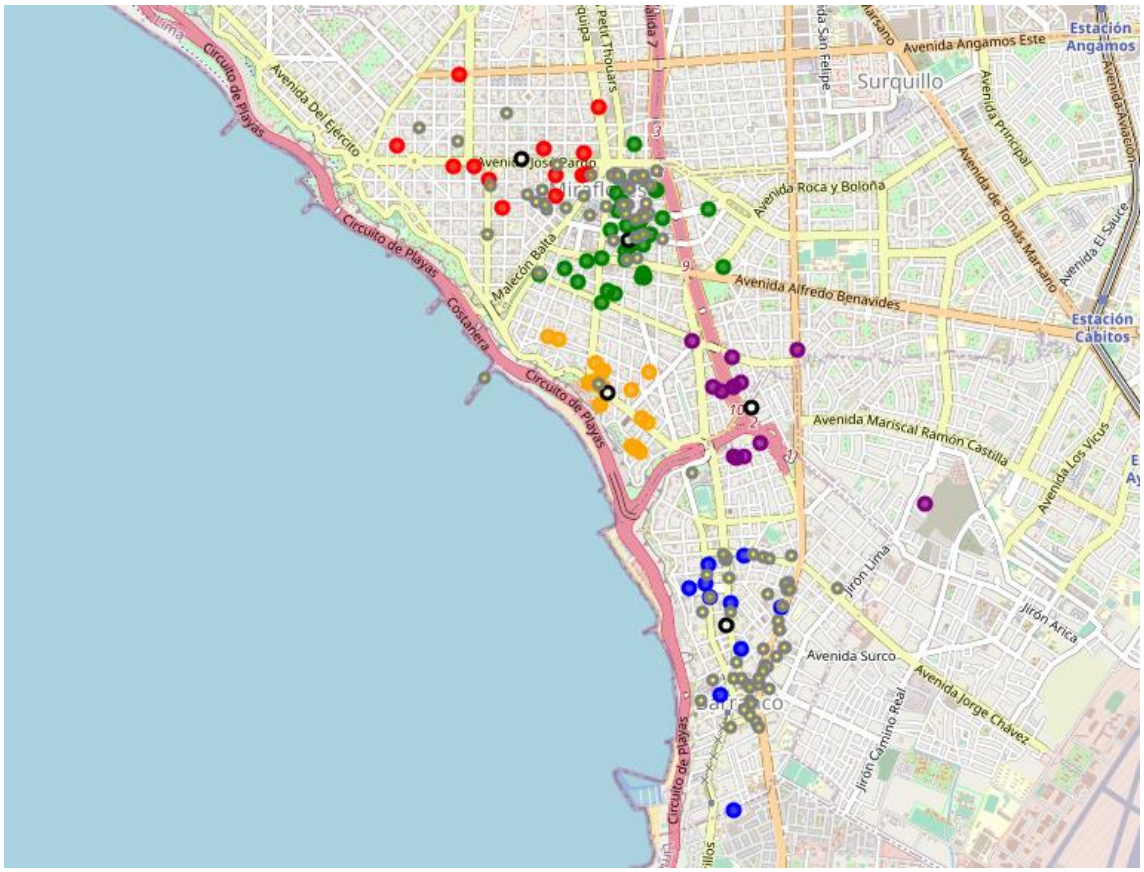
Also, we will set a new dataframe with each cluster center

	K_clusters	Latitude	Longitude	cluster_size
0	0	-12.132989	-77.020616	13
1	1	-12.118896	-77.033893	13
2	2	-12.145327	-77.022052	10
3	3	-12.123489	-77.027764	30
4	4	-12.132156	-77.028925	25

Now we can visualize the clusters with its centers of the hotels located before in Miraflores and Barranco districts.



To complete the information, we will add and overlap to the map the bars near the hotels.



4 Results

We have 5 clusters formed:

Cluster labeled 0 (purple cluster) with 13 hotels.
Cluster labeled 1 (red cluster) with 13 hotels.
Cluster labeled 2 (blue cluster) with 10 hotels.
Cluster labeled 3 (green cluster) with 30 hotels.
Cluster labeled 4 (orange cluster) with 25 hotels.

We also add the location of the bars for both districts.

5 Discussion

In the districts of Miraflores and Barranco you can see the formation of 5 well-defined clusters. This is due to the places of interest that exist. For example, between the red cluster and the green cluster, there is the Parque

Central de Miraflores, it is one of the most popular and visited parks in Lima surrounded by several commercial establishments, crafts, bars, and restaurants.

Similar happens in Barranco, where you can see the blue cluster. There is the Parque Municipal de Barranco, an old park dating from 1898, there and surrounding are cultural activities and there are also restaurants and bars.

The orange cluster and the red cluster contain a great number of potential location candidates. The hotels in these clusters are close to commercial establishments, close to the sea view, and the area has a low density of bars nearby. For these reasons, these areas would be recommended for a more detailed analysis depending on the bar specifications to be implemented.

6 Conclusion

The location of hotels and bars around tourist attractions is notorious.
The location of bars is not dependent on the location of the hotels.
For this capstone project, we found locations of high interest, even so, for more detailed and accurate reference, the data set can be expanded to the specific characteristics of the neighborhood or street.