IBM Data Science Course Capstone Project

Muhammed Ekrem Tecim

1. Problem Definition

A successful owner of medium and high-end restaurants(Nusret☺) decided to open a new restaurant in Barcelona. After visiting the city many times in recent years, he could not ignore the big bang in the field of gastronomy. Nusret is keen to open a new unit that will focus on Middle Eastearn meat-steak based cuisine. Given the price level at which the restaurant will operate, the aim is to find the most suitable place in an area where gastronomy is booming and is easily accessible for tourists and wealthier local citizens.

The assumption and business logic behind this analysis is by using unsupervised clustering of districts that can provide us the list of considirable restuarant. The intent is that the restaurant to be situated close to one of the gastronomical centres and touristic hotspots.

2.Data used for Analysis

To perform this analysis, data is needed on below:

List of the districts of Barcelona

Geo-coordinates of the districts in Barcelona

Top venues of districts

List of districts will be obtained from Wikipedia.
(https://en.wikipedia.org/wiki/Districts_of_Barcelona)

Geo-coordinates of districts will be obtained with the help of the geocoder tool in the notebook.

Top venues data will be obtained from Foursquare through an API.

3.Methodology

After tidying up and exploring the data, we will apply the K-means machine learning technique for creating clusters of districts. We will use the silhouette score for choosing the optimal number of clusters.

As part of preparing the data, we start by creating a list of districts in Barcelona and add the geo-coordinates of each district to this table. That is done by first importing a list of districts and then using this list and geocode python library, we add the latitude and longitude coordinates to

each district. After performing this task, we get the following table that we use in pandas dataframe format.

Now that we have the dataset ready, we perform clustering. For this, unsupervised machine learning technique will be used based on K-means. For K-means clustering, we need to decide on the number of clusters that we want to use. To avoid the trial and error approach, the silhouette score was used. Later we found optimal number of cluster by using these scores.

3.Limitation

-The analysis is performed on 10 districts in Barcelona.

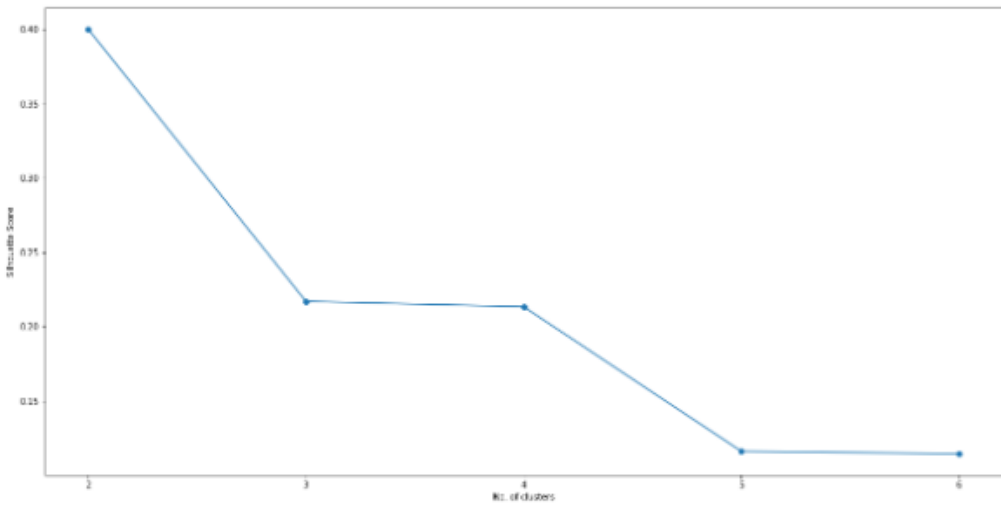
- The analysis is performed on a district level.

-When collecting venues a 500 meter radius is used around the centre coordinates of the districts. The number of collected venues is limited to 50 per districts

4.Results

Clusters

We cluster the districts based on kmeans algortihm. We see thar 2 cluster is giving us best result. Also it can be said that the clustering based on district can result less reliable based on performance scores.

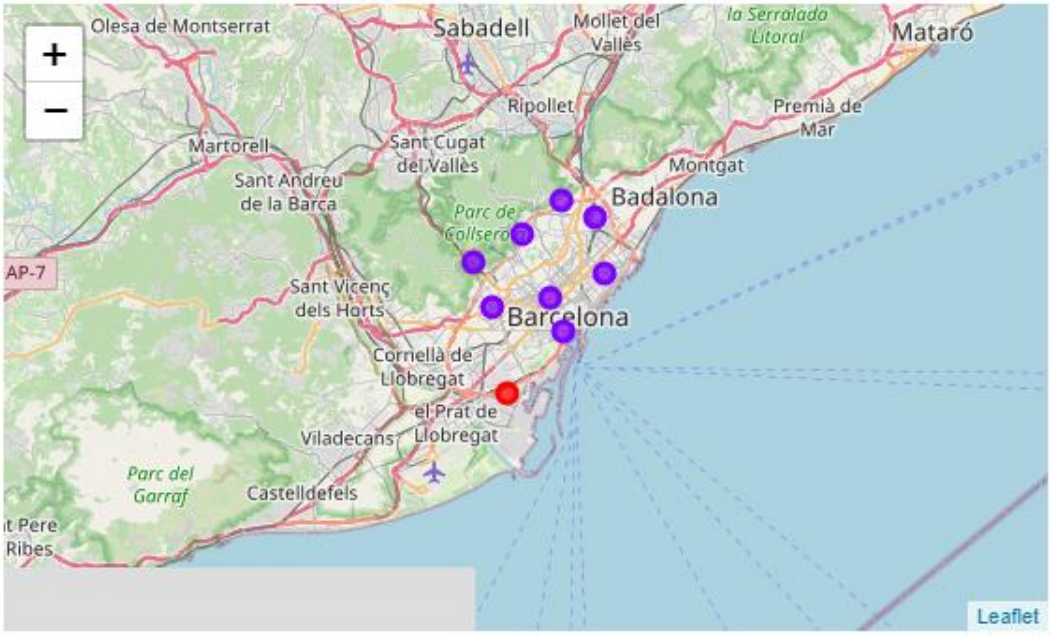


[152]:

```
KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
    n_clusters=2, n_init=10, n_jobs=None, precompute_distances='auto',
    random_state=0, tol=0.0001, verbose=0)
```

[153]:

[153]:



Cluster 1

]: `s_merged.loc[s_merged['Cluster Labels'] == 0, s_merged.columns[[0] + list(range(5, s_merged.shape[1]))]]`

]:

	Number	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
2	3	0.0	History Museum	Clothing Store	Restaurant	Farmers Market	Convenience Store	Cosmetics Shop	Country Dance Club	Cultural Center	Dessert Shop	Diner

Cluster 2

]: `s_merged.loc[s_merged['Cluster Labels'] == 1, s_merged.columns[[0] + list(range(5, s_merged.shape[1]))]]`

]:

	Number	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	1	1.0	Cocktail Bar	Hotel	Plaza	Boat or Ferry	Bar	Tapas Restaurant	Pizza Place	Spanish Restaurant	Art Gallery	Theater
1	2	1.0	Hotel	Mediterranean Restaurant	Tapas Restaurant	Bakery	Hostel	Bookstore	Cocktail Bar	Boutique	Cosmetics Shop	Spanish Restaurant
3	4	1.0	Hotel	Restaurant	Spanish Restaurant	Bakery	Park	Garden	Italian Restaurant	Diner	Bar	Tram Station
4	5	1.0	Light Rail Station	Convenience Store	National Park	Ice Cream Shop	Building	Plaza	BBQ Joint	Farmers Market	Falafel Restaurant	Exhibit
6	7	1.0	Park	Chinese Restaurant	Spanish Restaurant	Grocery Store	Farmers Market	Farm	Outdoor Sculpture	Plaza	Basketball Court	Soccer Field
7	8	1.0	Baby Store	Castle	Skate Park	Metro Station	Spanish Restaurant	Fried Chicken Joint	Falafel Restaurant	Gym	Cultural Center	Dessert Shop
8	9	1.0	Clothing Store	Cosmetics Shop	Tapas Restaurant	Burger Joint	Spanish Restaurant	Sandwich Place	Café	Fast Food Restaurant	Electronics Store	Women's Store
9	10	1.0	Mediterranean Restaurant	Pizza Place	Bakery	Italian Restaurant	Park	Performing Arts Venue	Plaza	Restaurant	Falafel Restaurant	Brewery

5. Discussion and Recommendations

We suggest Nusret to open sreak in first cluster due the common places of these districts are has diffirent type of restaurants and the cluster has the diversity.

6s. Conclusion

This article discussed the process of finding an answer, although hypothetical, for a business problem like real life. The analysis was based on the toolset of data science and was largely based on the use of Python and Python libraries such as Pandas, Scikit, Folium. The output of the analysis provided a comprehensive basis for the proposal for the business issue in question.