CAPSTONE PROJECT

IBM DATA SCIENCE PROFESSIONAL CERTIFICATE

MAY 22, 2021

Rafael Cidade da Silveira

Segmenting and Clustering Neighborhoods in Dubai and Doha

1. Introduction

This article is the second part of my capstone project for [Coursera IBM Data Science Professional Certificate](), a 10-course program offerd by IBM that explores several disciplines of the Data Science field. Some of those disciplines were applied to this project, wich is based on **Python**.

1.1. Background

Dubai and **Doha** are cities located on the coast of the Persian Gulf, and they are very similar to each other. Not only because of their geographic location, climate and futuristic skyscrapers, but also for both being ranked among the most high tech and safest cities in the world.

Doha and Dubai are situated on the coast of the Arabian Peninsula (Google Maps)

Dubai is the most populous city in the United Arab Emirates (UAE) and the capital of the Emirate of Dubai. The population of Dubai is estimated at around 3,400,800 and its Economy represents a gross domestic product (GDP) of US$102.67 billion.

Doha is the capital and most populous city of Qatar. Its population is estimated at around 2,382,000. In terms of Economy, Doha's GDP is around US$146.09 billion.

The life cost and average salary in both cities are similar too. According to the website [Livingcost.org](https://livingcost.org), the cost of living in Doha is 2% less expensive than in Dubai.

	Doha	Dubai
Cost of living One person	$1804	$1836
Cost of living Family	$3984	$4258
One person rent	$1132	$1108
Family rent	$1970	$2059
Food Expenses	$447	$433
Transport Expenses	$94.1	$150
Monthly salary after tax	$3096	$2891
Population	1.31M	2.5M

Doha vs Dubai comparison (https://livingcost.org)

1.2. Problem description

Dubai and Doha are also very similar in terms of opportunities for work, especially in IT. Based on that, we may face the following situation: an IT professional, based in Dubai, who got a great job offer from a big company in Doha, decides to move to Qatar's capital.

In Doha, the professional would like not only to live near the new job but also to settle in somewhere similar to where he or she is currently based. Naturally, if we are used to live near places that make our lives easier and more confortable (such as gyms, restaurants, groceries), we'll look for neighborhoods with the same characteristics when moving to a new city.

So, the challenge here is to use Data Science to help that professional in finding the best neighborhood.

2. Data

In this project, I basically used the following data: **list of neighborhoods in Dubai and Doha** (containing names and geolocation of each one of them), and **Foursquare data** (list of venues) about each one of those neighborhoods.

The source of the list of neighborhoods, in Doha and in Dubai, is Wikipedia. There are specific pages that list neighborhoods of both cities, with basic information, such as name and population, and a link to the neighborhood wiki page, where is informed its geolocation.

Foursquare is a technology company, and one of its products is **Foursquare City Guide**, an app that provides recommendations of places to go near a location. Foursquare's API is the source of the main neighborhood data that I used in the project.

3. Methodology

Using Foursquare API and some Data Science tools and techniques, I segmented and compared neighborhoods of the two cities.

The first step consisted in getting neighborhood information from Wikipedia. To do that, I used a technique widely known as *web scraping,* that involves reading the page's source code to extract data from it.

V·T·E	Neighbourhoods and communities in Dubai [hide]
Deira and eastern Dubai	Abu Hail · Al Baraha · Al Buteen · Al Dhagaya · Al Garhoud · Al Hamriya Port · Al Karama · Al Khabisi · Al Mamzar · Al Mizhar · Al Muraqqabat · Al Murar · Al Muteena · Al Nahda · Al Qusais · Al Ras · Al Rashidiya · Al Rigga · Al Sabkha · Al Twar · Al Waheda · Al Warqaa · Ayal Nasir · Dubai International Airport · Hor Al Anz · Mirdif · Muhaisnah · Nad Al Hammar · Nad Shamma · Naif · Port Saeed · Rigga Al Buteen · Umm Ramool · Warisan · Al Amardhi
Bur Dubai and western Dubai	Al Bada · Al Barsha · Al Hamriya · Al Hudaiba · Al Jaddaf · Al Jafilia · Al Karama · Al Kefaf · Al Manara · Al Mankhool · Al Markada · Al Quoz · Al Rifa · Al Safa · Al Satwa · Al Shindagha · Al Souk Al Kabir · Al Sufouh · Al Wasl · Bu Kadra · Business Bay · Dubai Marina · Emirates Hills · Downtown Dubai · Dubai International City · Jebel Ali · Jumeirah · Jumeirah Islands · Jumeirah Lake Towers · Nad Al Sheba · Oud Metha · Port Rashid · Ras Al Khor · Ras Al Khor Industrial Area · Trade Centre 1 · Trade Centre 2 · Umm Al Sheif · Umm Hurair · Umm Suqeim · Zabeel

Neighborhoods section — Dubai Communities Wikipedia page

V·T·E	Neighbourhoods and communities in Doha [hide]
Census-designated districts	Al Bidda · Al Dafna · Ad Dawhah al Jadidah · Al Egla · Al Hilal · Al Jasrah · Al Kharayej · Al Khulaifat · Al Mansoura · Al Markhiya · Al Messila · Al Mirqab · Al Najada · Al Qassar · Al Rufaa · Al Sadd · Al Souq · Al Tarfa · Al Thumama · Barahat Al Jufairi · Dahl Al Hamam · Doha International Airport · Doha Port · Duhail · Fereej Abdel Aziz · Fereej Al Asmakh · Fereej Al Nasr · Fereej Bin Durham · Fereej Bin Mahmoud · Fereej Bin Omran · Fereej Kulaib · Fereej Mohammed Bin Jasim · Hamad Medical City · Hazm Al Markhiya · Industrial Area · Jabal Thuaileb · Jelaiah · Jeryan Nejaima · Lejbailat · Lekhwair · Leqtaifiya · Madinat Khalifa North · Madinat Khalifa South · Musheireb · Najma · New Al Hitmi · New Al Mirqab · New Salata · Nuaija · Old Airport · Old Al Ghanim · Old Al Hitmi · Old Salata · Onaiza · Ras Abu Aboud · Ras Abu Fontas · Rawdat Al Khail · Rumeilah · Umm Ghuwailina · Umm Lekhba · Wadi Al Banat · Wadi Al Sail · West Bay
	See also: *Zones of Qatar*
V·T·E	Municipality of Ad-Dawhah topics [show]

Neighborhoods section — Doha Communities Wikipedia page

I imported the libraries *request*, to download the source code of the urls, and *BeautifulSoap*, to handle the HTML code and extract information from it. The result was stored in a *pandas* data frame.

```
In [2]: #create a dataframe to store neighborhoods of both cities
neighborhood_data = pd.DataFrame(columns=["City", "Neighborhood", "Latitude", "Longitude"])

# webscrape neighborhoods from Doha wikipedia page
r = requests.get('https://en.wikipedia.org/wiki/List_of_communities_in_Doha')

soup = BeautifulSoup(r.text.replace('\n', ''), "html.parser") #replaces line break

#finds the correct table based on its class
doha_neighborhood_table = soup.find("table", {"class": "wikitable"})

for row in doha_neighborhood_table.find("tbody").find_all("tr"):
    if not row.find_all("th"): #handle data only if no table head is found
        col = row.find_all("td")

        links = col[0].find_all("a", href=True)

        for link in links:
            neighborhood = link.text

            r = requests.get('https://en.wikipedia.org' + link["href"])
            coordinates = BeautifulSoup(r.text.replace('\n', ''), "html.parser").find("span", {"class": "geo-de
            latitude = coordinates[0].replace("°N","")
            longitude = coordinates[1].replace("°E","")

            neighborhood_data = neighborhood_data.append({"City":"Doha",
                                                          "Neighborhood":neighborhood,
                                                          "Latitude":float(latitude),
                                                          "Longitude":float(longitude)}, ignore_index=True)


neighborhood_data.head()
```

Out[2]:

	City	Neighborhood	Latitude	Longitude
0	Doha	Al Bidda	25.29972	51.51972
1	Doha	Al Dafna	25.32389	51.53056
2	Doha	Ad Dawhah al Jadidah	25.27583	51.53361
3	Doha	Al Egla	25.38900	51.50950
4	Doha	Al Hilal	25.28667	51.53333

Having the neighborhoods coordinates, I defined a function to call Foursquare API for each location and get the nearby venues, within a 1000-meter radius of that location.

The results were stored in another data frame, merging the previous information about the neighborhoods and the data of each venue returned from Foursquare API.

```
nearby_venues.columns = ['City', 'Neighborhood',
                 'Neighborhood Latitude',
                 'Neighborhood Longitude',
                 'Venue',
                 'Venue Latitude',
                 'Venue Longitude',
                 'Venue Category']
```

In order to apply **Machine Learning** algorithms for clustering, I converted the categorical variable “Venue Category” into dummy/indicator variables, resulting in a data frame with 312 columns (one for each different category).

In [13]:
```
#create a new dataframe, converting categories into indicator variables
# one hot encoding
doha_dubai_onehot = pd.get_dummies(doha_dubai_venues[['Venue Category']], prefix="", prefix_sep="")

# add city and neighborhood columns to dataframe
doha_dubai_onehot['Neighborhood'] = doha_dubai_venues['Neighborhood']
doha_dubai_onehot['City'] = doha_dubai_venues['City']

doha_dubai_onehot.head()
```

Out[13]:

	Accessories Store	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Arcade	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	R
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

In [14]:
```
doha_dubai_onehot.shape
```

Out[14]: (7034, 312)

With *numpy*, another useful library, I arranged the most frequent venues of each neighborhood and sorted the results in descending order. For the first time I was able to visually compare the neighborhoods.

	City	Neighborhood	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	Doha	Ad Dawhah al Jadidah	Hotel	Indian Restaurant	Café	Fast Food Restaurant	BBQ Joint	Middle Eastern Restaurant	Lounge	Jewelry Store	Burger Joint	Asian Restaurant
1	Doha	Al Bidda	Park	Bowling Alley	Intersection	Theater	Trail	Boat or Ferry	Beach	Harbor / Marina	Historic Site	Seafood Restaurant
2	Doha	Al Dafna	Hotel	Coffee Shop	Café	Restaurant	Italian Restaurant	Lebanese Restaurant	Lounge	Spa	Bar	Steakhouse
3	Doha	Al Hilal	Café	Hotel	Middle Eastern Restaurant	Coffee Shop	BBQ Joint	Harbor / Marina	Restaurant	Museum	Fried Chicken Joint	Indian Restaurant
4	Doha	Al Jasrah	Hotel	Café	Middle Eastern Restaurant	Coffee Shop	BBQ Joint	Restaurant	Indian Restaurant	Museum	Turkish Restaurant	Italian Restaurant

After that, I imported *KMeans* to finally cluster the neighborhoods based on their similarities.

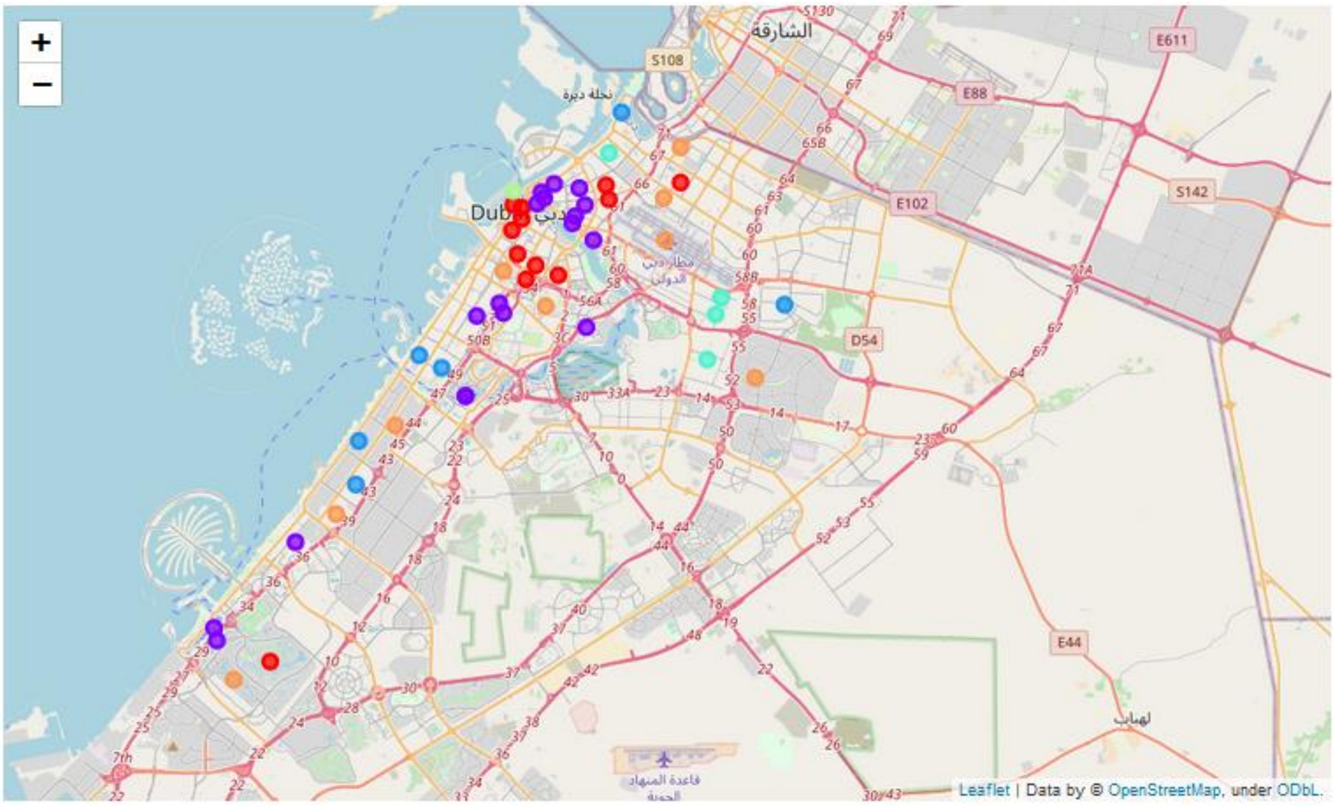
Using *silhouette_score*, from *sklearn.metrics*, I found out that **6** was the best number of clusters, based on the neighborhood data I had.

With KMeans method, the neighborhoods were labeled with values ranging from 0 to 5.

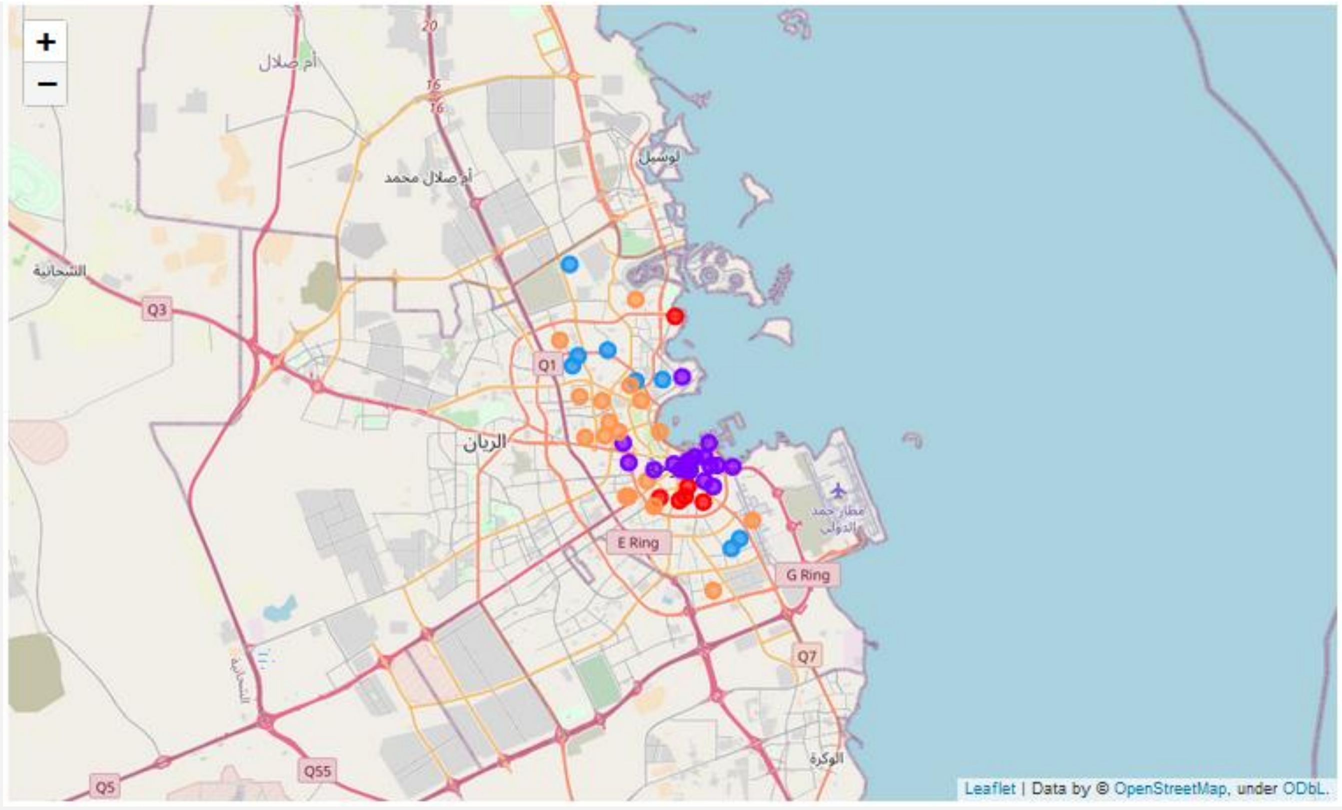
```
In [20]: # import k-means from clustering stage
         from sklearn.cluster import KMeans

         kclusters = 6
         df = doha_dubai_grouped.drop(['City', 'Neighborhood'], 1)

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

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

Out[20]: array([5, 4, 0, 0, 0, 0, 5, 1, 4, 0, 0, 5, 0, 0, 0, 4, 0, 1, 4, 0, 0, 4,
                5, 4, 4, 4, 0, 0, 1, 1, 4, 4, 4, 1, 4, 0, 5, 4, 4, 4, 1, 1, 0, 0,
                1, 5, 0, 4, 0, 4, 4, 1, 1, 0, 0, 5, 0, 0, 5, 0, 0, 4, 5, 5, 5, 1,
                1, 5, 0, 0, 0, 4, 5, 5, 2, 5, 0, 0, 4, 0, 5, 5, 0, 4, 2, 4, 1, 0,
                0, 0, 4, 0, 5, 5, 1, 4, 0, 1, 2, 2, 0, 5, 3, 0, 0, 0, 0, 4, 1, 1,
                4], dtype=int32)
```

4. Results

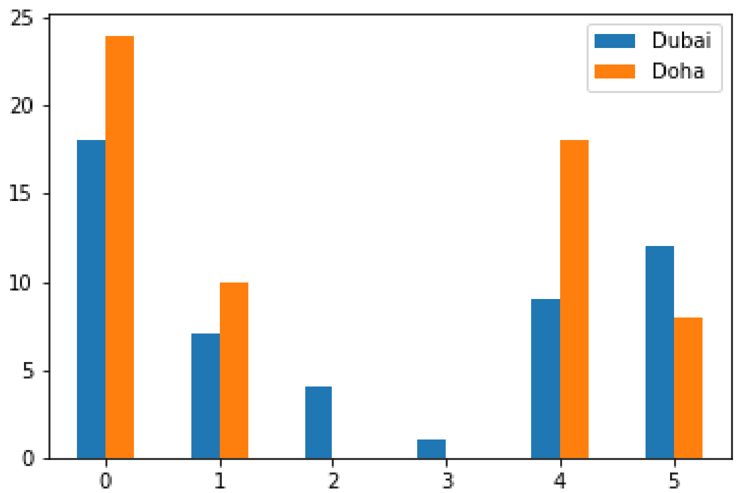
I used the library *folium* to create maps of Dubai and Doha, marking neighborhoods with six different colors, one for each cluster.



Map of Dubai with clustered neighborhoods



Map of Doha with clustered neighborhoods



Number of neighborhoods of Dubai and Doha in each cluster

I also grouped, by cluster, the most common venues of neighborhoods.

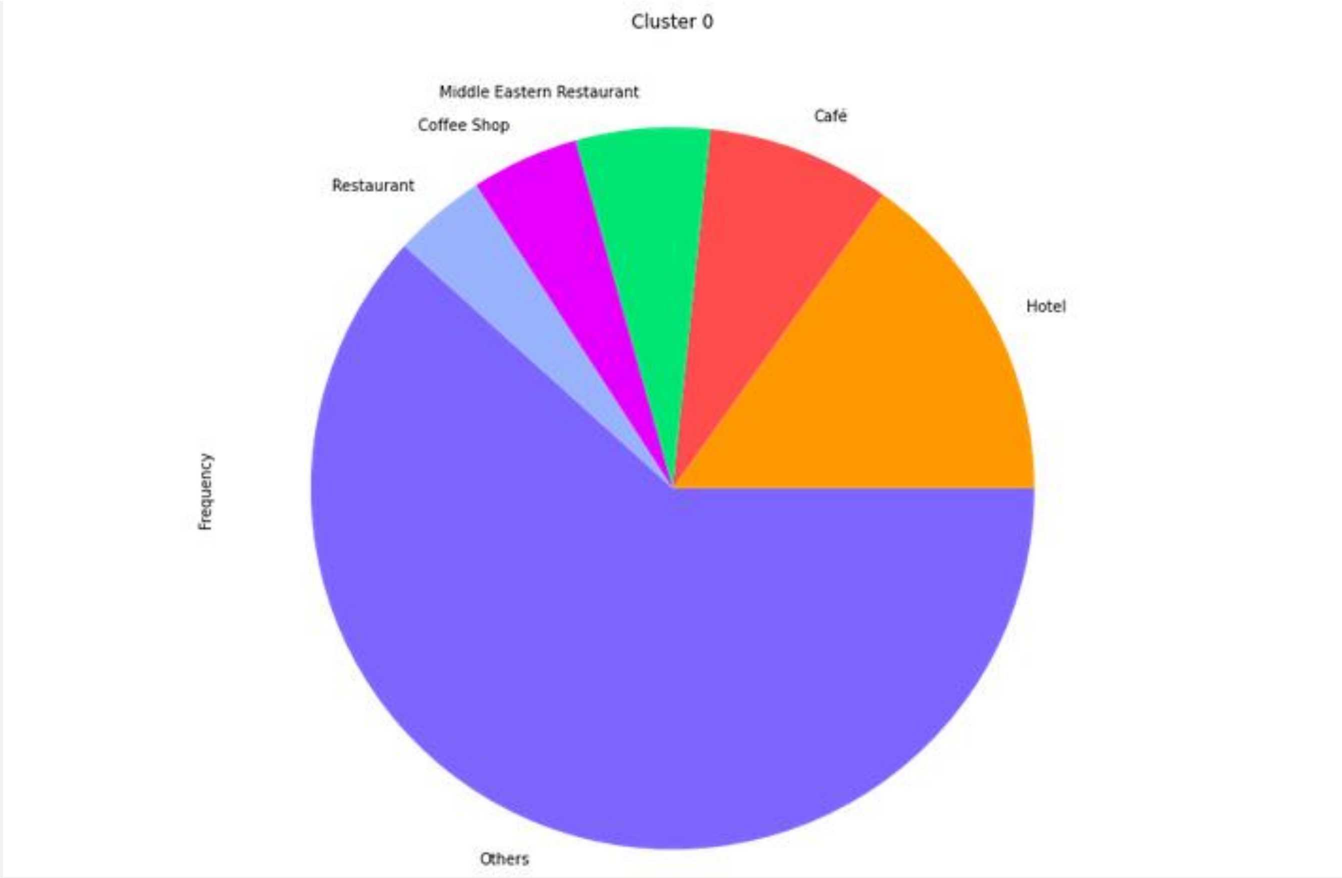
In [35]:
```
doha_dubai_grouped_cluster = doha_dubai_grouped.groupby(['Cluster Labels']).mean().reset_index()
doha_dubai_grouped_cluster.head(6)
```

Out[35]:

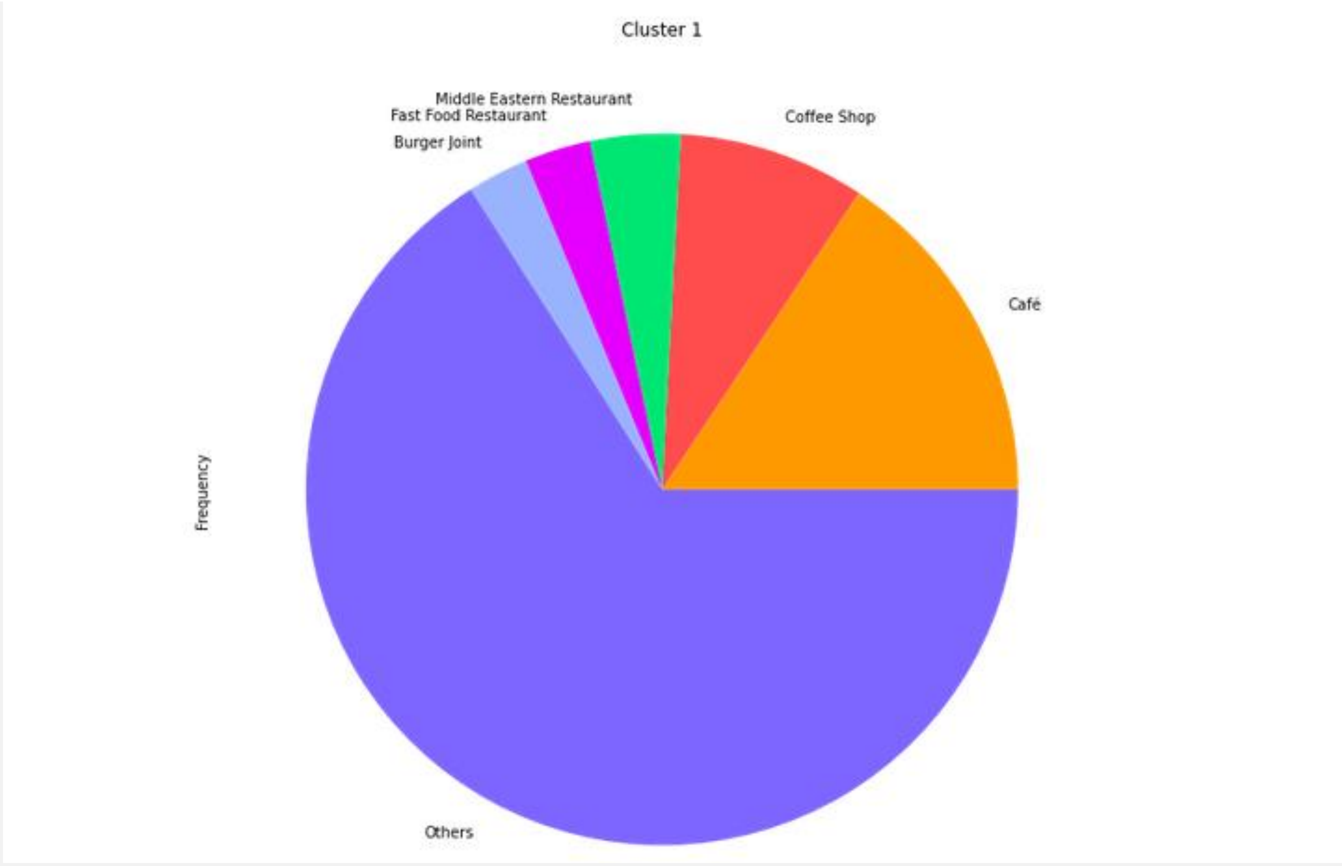
	Cluster Labels	Accessories Store	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Arcade	Argentinian Restaurant	Ar Gallery
0	0	0.000000	0.001615	0.000722	0.000000	0.000000	0.000000	0.000000	0.000000	0.005876	0.000960	0.000238	0.00461
1	1	0.000000	0.001314	0.001814	0.000000	0.000000	0.000000	0.000000	0.000000	0.004977	0.003438	0.000000	0.00000
2	2	0.013889	0.000000	0.000000	0.013889	0.000000	0.000000	0.000000	0.007576	0.000000	0.000000	0.000000	0.00000
3	3	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.041667
4	4	0.003419	0.000828	0.003662	0.001538	0.000588	0.009701	0.002979	0.003687	0.021811	0.000712	0.000000	0.00095
5	5	0.002473	0.000500	0.002704	0.000000	0.000000	0.000000	0.000000	0.000000	0.005197	0.001564	0.000000	0.005327

Then, I could plot pie charts of each cluster and get a better visualization of them:

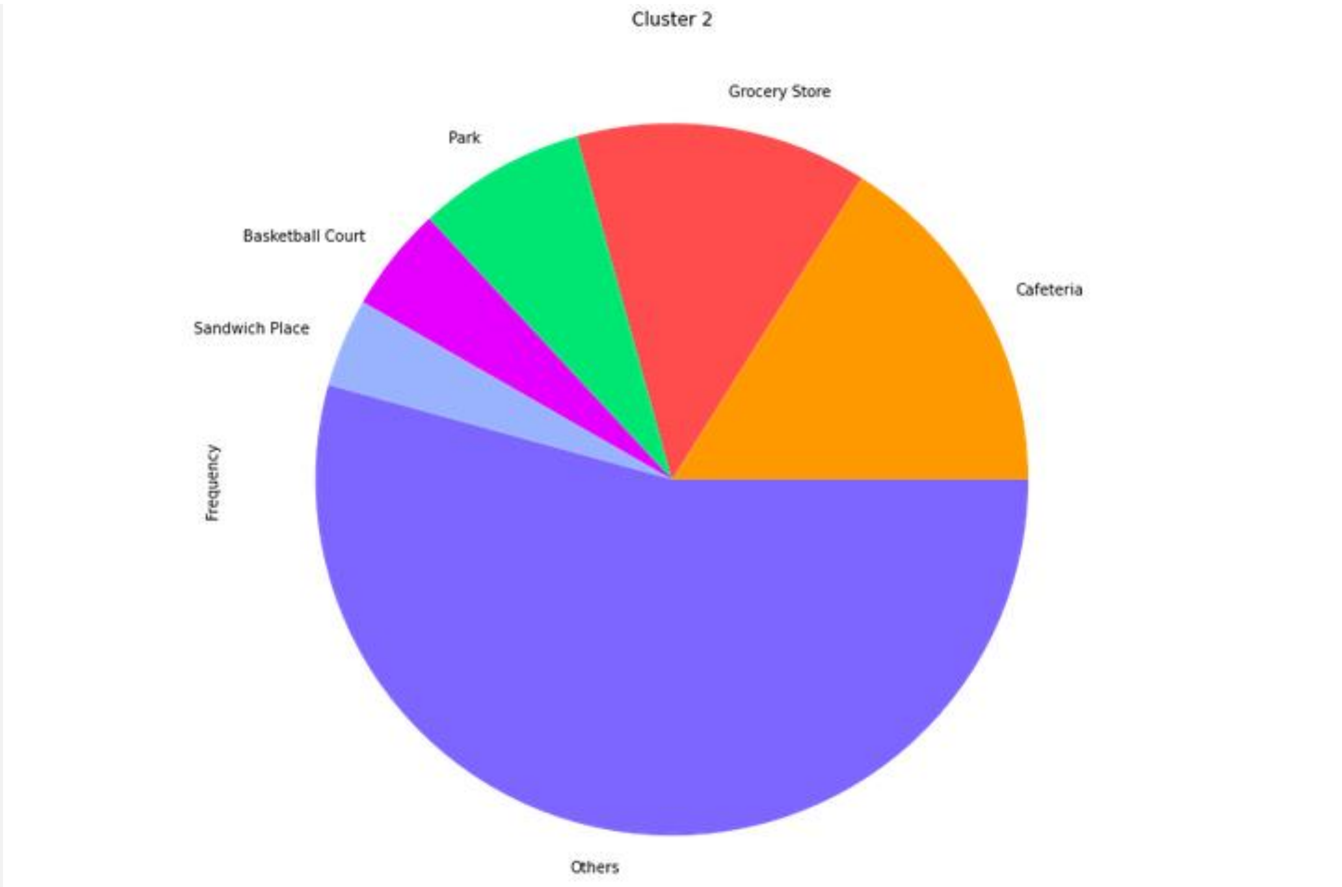
Cluster 0



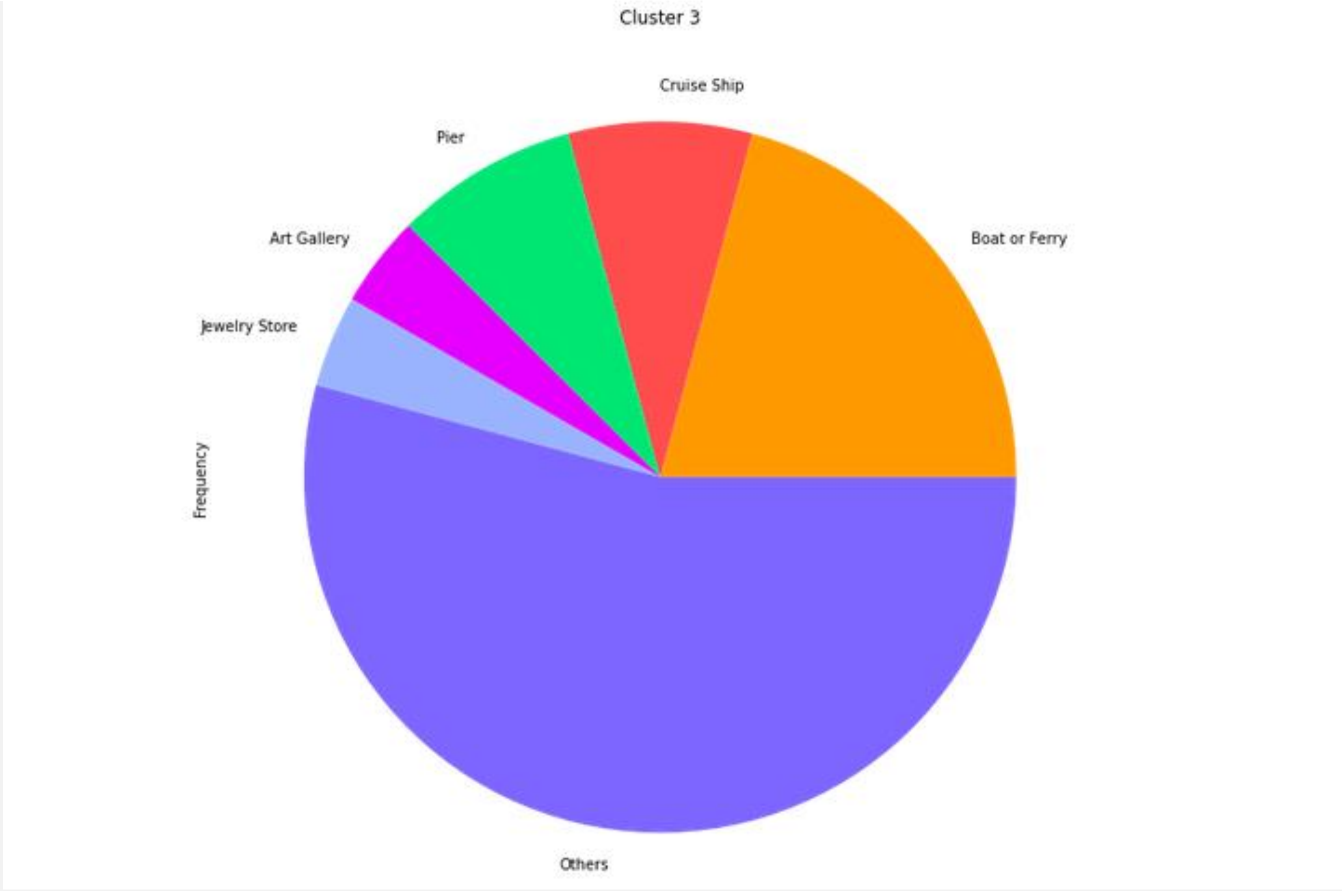
Cluster 1


Cluster 1
Middle Eastern Restaurant
Fast Food Restaurant
Burger Joint
Coffee Shop
Café
Frequency
Others

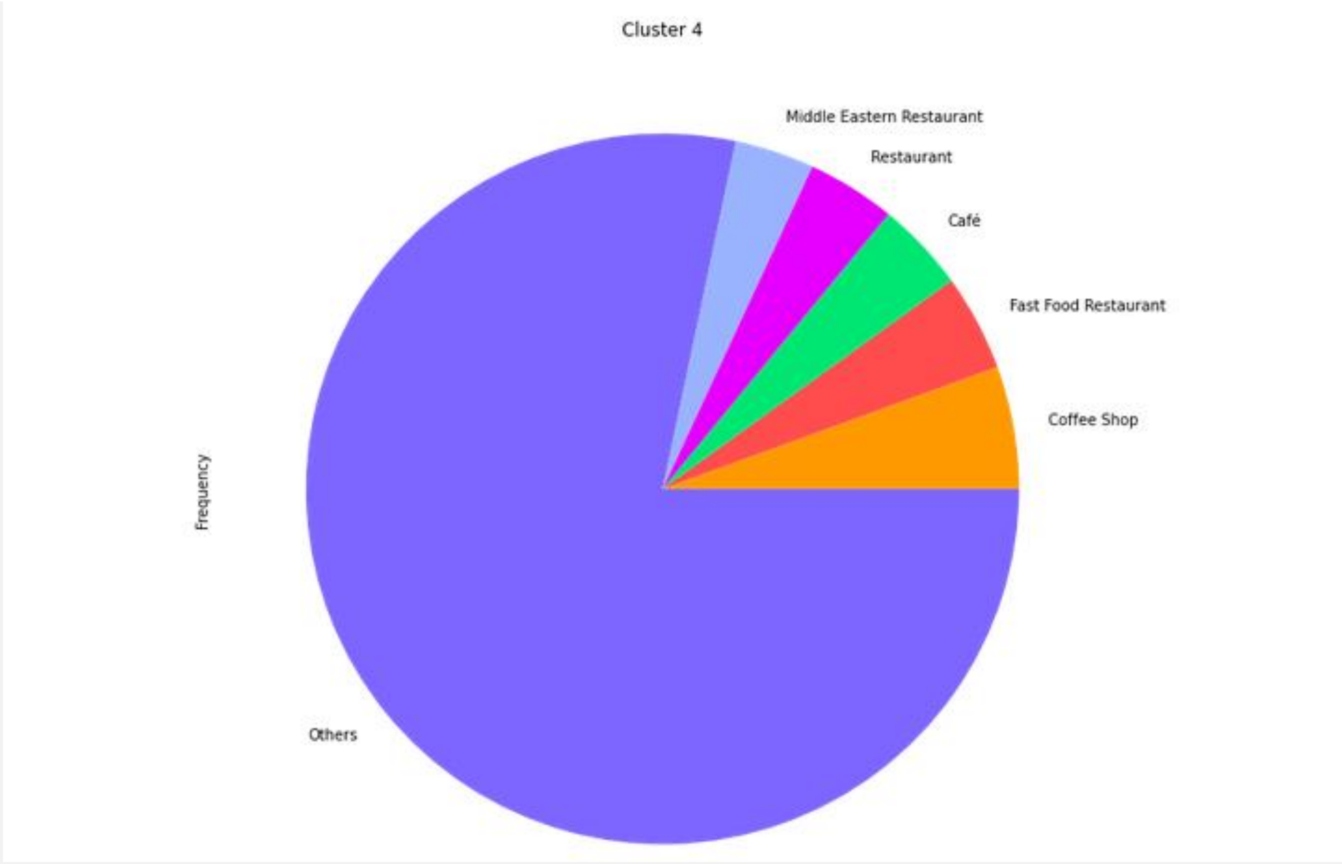
Cluster 2


Cluster 2
Grocery Store
Park
Basketball Court
Cafeteria
Sandwich Place
Frequency
Others

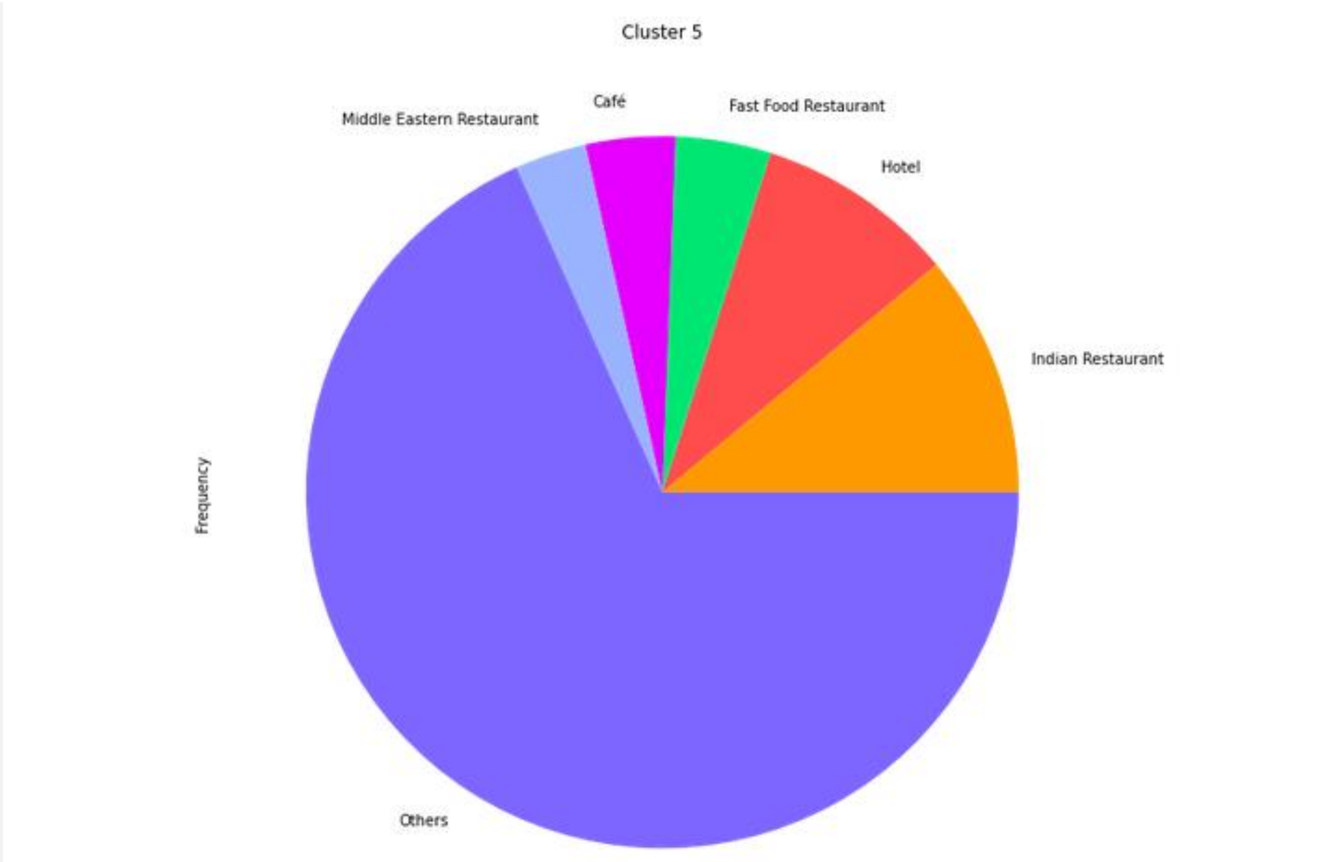
Cluster 3



Cluster 4


Cluster 4
Middle Eastern Restaurant
Restaurant
Café
Fast Food Restaurant
Coffee Shop
Frequency
Others

Cluster 5


Cluster 5
Middle Eastern Restaurant
Café
Fast Food Restaurant
Hotel
Indian Restaurant
Frequency
Others

5. Discussion

Analyzing the neighborhoods of cluster 0, I could observe that **Hotel** is the most common venue category in many of them. It shows us how important is the definition of the best *k*, to cluster data in the most proper way.

	Cluster Labels	City	Neighborhood	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 Mos Commo Venu
2	0	Doha	Al Dafna	Hotel	Coffee Shop	Café	Restaurant	Italian Restaurant	Lebanese Restaurant	Lounge	Spa	Ba
3	0	Doha	Al Hilal	Café	Hotel	Middle Eastern Restaurant	Coffee Shop	BBQ Joint	Harbor / Marina	Restaurant	Museum	Fried Chicke Joir
4	0	Doha	Al Jasrah	Hotel	Café	Middle Eastern Restaurant	Coffee Shop	BBQ Joint	Restaurant	Indian Restaurant	Museum	Turkis Restaurar
5	0	Doha	Al Khulaifat	Hotel	Indian Restaurant	Café	Athletics & Sports	Restaurant	Middle Eastern Restaurant	Fast Food Restaurant	Beach	Nightclu
9	0	Doha	Al Mirqab	Hotel	Café	Middle Eastern Restaurant	Restaurant	Coffee Shop	Museum	Hookah Bar	Mediterranean Restaurant	Flea Marke
10	0	Doha	Al Najada	Hotel	Café	Middle Eastern Restaurant	Coffee Shop	Restaurant	Bakery	BBQ Joint	Seafood Restaurant	Fried Chicke Joir
12	0	Doha	Al Rufaa	Hotel	Café	Indian Restaurant	Middle Eastern Restaurant	Restaurant	Museum	Fast Food Restaurant	BBQ Joint	Coffee Sho
13	0	Doha	Al Sadd	Hotel	Italian Restaurant	Café	Coffee Shop	Middle Eastern Restaurant	Nightclub	Thai Restaurant	Bar	Lebanes Restaurar

Another demonstration of the best k importance, is related to Cluster 3. It has only one single neighborhood, Port Rashid, in Dubai. That happend because of Port Rashid's singular characteristics. The most common venues of the neighborhood are related to Boat or Ferry, Cruise Ship, Pier and Port.

If I had opted for a lower *k*, it is possible that Port Rashid would be clustered with neighborhoods much differents of it.

In [30]:
```
cluster3 = doha_dubai_merged.loc[doha_dubai_merged['Cluster Labels'] == 3]
cluster3
```

Out[30]:

	Cluster Labels	City	Neighborhood	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	Latitu
102	3	Dubai	Port Rashid	Boat or Ferry	Cruise Ship	Pier	Port	Tunnel	Shopping Mall	Museum	Flower Shop	Middle Eastern Restaurant	Bed & Breakfast	25.274

6. Conclusion

The project achieved its purpose and delivered segmented neighborhoods of Dubai and Doha, and detailing the most common venues and their frequency in neighborhoods and groups of neighborhoods.

It is a useful study for those who want to find a new neighborhood to live in one of the cities or simply for curious people and data science enthusiasts.

The Python notebook of this project can be checked here.

7. References

https://pandas.pydata.org/docs/user_guide/index.html

https://developer.foursquare.com/docs/

https://en.wikipedia.org/wiki/List_of_communities_in_Dubai

https://en.wikipedia.org/wiki/List_of_communities_in_Doha

https://livingcost.org/cost/doha/dubai

https://www.businessinsider.com/most-innovative-cities-in-the-world-in-2018-2018-11