



IBM APPLIED DATA SCIENCE CAPSTONE FINAL PROJECT

Optimal Location for a Healthy Food
Brand in Doha, Qatar

Contents



INTRODUCTION :
UNDERSTANDING THE
PROBLEM



DATA COLLECTION,
CLEANING AND
PREPARATION



METHODOLOGY



RESULTS & CONCLUSION

Introduction

“I have a home-run healthy snacks company and I am looking for potential spot to set up shop here in Doha, Qatar”

- The idea of healthy eating is gaining popularity here in Qatar
- it is important for people to easily find a healthy option near to wherever they are
- In this project we want to focus on localities that are popular in Doha
- Additionally, it would also be beneficial to be near gyms/fitness centers
- *And near schools*

Data

“We will focus this research on the zones under the Doha municipality”

- 1) Get the names and zones from Wikipedia
- 2) Clean, optimize and prepare data set
- 3) Get latitude and longitude of the zones from latlong.net
- 4) Using Foursquare, identify the popular spots in each zone
- 5) Implement Kmeans on the data to form clusters
- 6) Analyze clusters

Methodology

- Use **Beautiful Soup** to scrape tables off of Wikipedia
- Data was cleaned using **pandas**' built-in commands
- We populate the Dataframe with the latitude and longitude obtained through the website → <https://www.latlong.net/>
- use the **Folium package** to plot the zones on the map
- use **Foursquare API** to get data for the popular spots in our zones
- use the **KMeans** module from **sklearn package** to cluster the data



Folium



API



Methodology

Why KMeans?

The K-means clustering algorithm uses iterative refinement to produce a final result. The algorithm inputs are the number of clusters K and the data set. The data set is a collection of features for each data point.

Once we have grouped the data into K clusters, we can analyze the results to select the best option for our purpose

Why not DBSCAN?

K-means is intended to find K clusters on a dataset based on distance to center of the clusters. DBSCAN is density-based algorithm so concept of distance is absent. The algorithm uses concept of reachability instead. Like, how many neighbours have a point within a radius. Depending on this idea, clusters are generated. In DBSCAN, this radius is fixed.

Working with real data over maps, it was thought K-means would be the better choice

Results

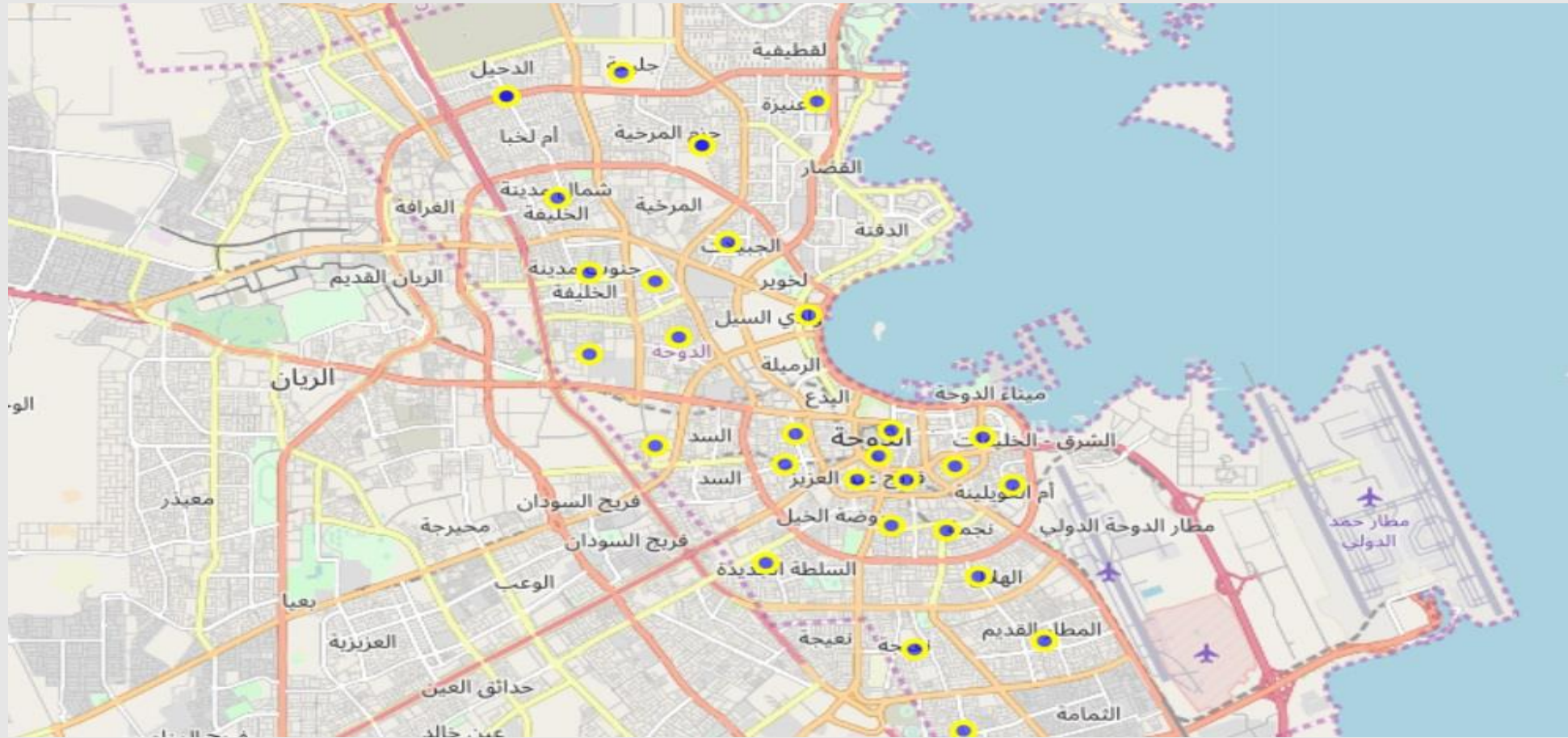
After scraping the data from Wikipedia page, cleaning it, optimizing it for use and populating the latitude and longitude column we are left with a database of about 30 zones

Out[4]:

	Zone	Districts	Population	Latitude	Longitude
0	3	Fereej Mohamed Bin Jasim	4886	25.2865	51.5296
1	4	Mushayrib	28069	25.2818	51.5275
2	14	Fereej Abdel Aziz	15706	25.2777	51.5242
3	15	Ad Dawhah al Jadidah	15920	25.2776	51.5321
4	16	Old Al Ghanim	16334	25.2800	51.5400
5	17	Al Rufaa	6026	25.2853	51.5444
6	22	Fereej Bin Mahmoud	28327	25.2803	51.5124
7	24	Rawdat Al Khail	18200	25.2860	51.5142
8	25	Fereej Bin Durham	37082	25.2693	51.5295
9	26	Najma	28228	25.2683	51.5387
10	27	Umm Ghuwailina	33262	25.2766	51.5492
11	30	Duhail	7705	25.3477	51.4675
12	31	Umm Lekhba	11897	25.3477	51.4675

Results

We then visualize these areas on the map



Results

Next, using Foursquare we get the popular venues around each of these areas. All in all we get 900 venues

```
print(Venues.shape)
Venues
```

(900, 7)

172]:

	Neighborhood	Neighborhood Latitude	Neighborhood
0	Fereej Mohamed Bin Jasim	25.2865	
1	Fereej Mohamed Bin Jasim	25.2865	
2	Fereej Mohamed Bin Jasim	25.2865	
3	Fereej Mohamed Bin Jasim	25.2865	
4	Fereej Mohamed Bin Jasim	25.2865	
5	Fereej Mohamed Bin Jasim	25.2865	
6	Fereej Mohamed Bin Jasim	25.2865	
7	Fereej Mohamed Bin Jasim	25.2865	
8	Fereej Mohamed Bin Jasim	25.2865	

Results

In each zone, I selected the top 5 categories and we started to get some promising results


```
temp = temp.round({'freq': 2})
print(temp.sort_values('freq', ascending=False))
print('\n')
```

Ad Dawhah al Jadidah

	venue	freq
0	Hotel	0.13
1	Mediterranean Restaurant	0.07
2	Café	0.07
3	Restaurant	0.07
4	Vegetarian / Vegan Restaurant	0.03

Al Dafna

	venue	freq
0	Coffee Shop	0.17
1	Café	0.17
2	American Restaurant	0.07
3	Park	0.07
4	Athletics & Sports	0.07



Al Hilal

	venue	freq
--	-------	------

Results

Finally, we divide the data into four clusters for further analysis:

Code	Cluster Labels	1st. Mos Common Venu
5124	0	Caf
5387	0	Caf
4675	0	Coffe Sho
5439	0	Caf
5176	0	Caf
5032	0	Caf

Code	Cluster Labels	1st. Mos Common Venu
95	1	
14	1	
53	1	Res
61	1	

Code	Cluster Labels	1st. Mos Common Venu
96	2	
44	2	
92	2	
92	2	
94	2	
34	2	
63	2	
92	2	

Code	Cluster Labels	1st. Mos Common Venu
5	3	
2	3	
1	3	
2	3	It Resta
8	3	
8	3	
4	3	

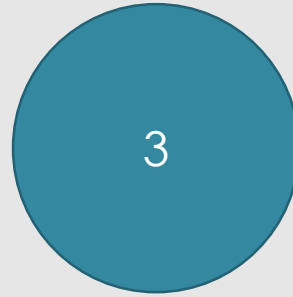
Discussion

Upon inspection of the clusters, it is seen that the most popular places across clusters is restaurants and cafes. Cluster 2 is too small so we will omit it. Readjusting our focus onto cluster 1,3 and 4



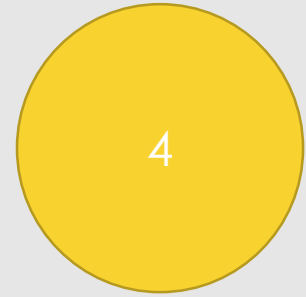
- densely populated
- has popular fitness and athletic spots
- has a park

Disadvantage: highly competitive area



- boasts a park
- has shopping mall
- has gym

Disadvantage: smallest of the three clusters



- most densely populated

Disadvantage: lacks gyms

Conclusion

Based on the information we have it would be best to do some more analyzing before making a decision. I would love to further check with of the 4 clusters has the most schools and it would also be beneficial to see traffic patterns around these areas, especially closer to noon and sunset.

For now, if I had to make a decision, I would open my healthy café in cluster 1 and take my chances with the competition 😊