# IBM Applied Data Science Capstone
# Final Project

Optimal Location for a Healthy Food Brand in Doha, Qatar

Contents

**Introduction**

This project is based on a real-life problem. I have a home-run healthy snacks company and I am looking for potential spot to set up shop here in Doha, Qatar. I'm going to try and be as practical as I can with this decision seeing how I am personally invested.

The idea of healthy eating is gaining popularity here in Qatar. Slowly but surely, people are making an effort to eat responsibly, but 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 we popular in Doha since that means most foot-traffic for our business and ease of access for clients.

Additionally, it would also beneficial to be near schools and gyms/fitness centers as the demographics that come to both these places are our target audience.

Using all my data science experience, I set forth to discover the best location possible for this investment.

**Data**

The most important factors to consider are:

* popular spots around Doha

Qatar is divided into eight municipalities, and each municipality is further divided into zones. We will focus this research on the zones under the *Doha municipality*, which will serve as our defined neighbourhoods.

*Step 1*

I will first get the names and zones of the neighbourhoods for the Doha Municipality from the Wikipedia page → https://en.wikipedia.org/wiki/Zones_of_Qatar

*Step 2*

The data needs to be cleaned

- We will only consider zones with a minimum population of 3000

- Zones that are repeated in the list need to be merged
- We will drop any zones that has incomplete data

*Step 3*

Next we will get the latitude and longitude of all the zones on the list from latlong.net

*Step 4*

Using Foursquare we will identify the popular spots in each zone

*Step 5*

We will implement Kmeans on the data to form clusters

*Step 6*

Analyzing the clusters will give us realistic choices of where we can set up shop

**Note:** Qatar is a small country and Doha is just one of the country's municipalities. We further scale this down to zones within Qatar and it is no surprise that the latitudes and longitudes found were identical down to 2 decimal places. It was only at the 3rd decimal place where we saw any differences

**Methodology**

1. Firstly we use **Beautiful Soup** to scrape tables off of Wikipedia and convert them into a Dataframe using **pandas**.

2. Data was cleaned using **pandas'** built-in commands such as SELECT (MIN), MERGE and DROP

3. We populate the Dataframe with the latitude and longitude obtained through the website → https://www.latlong.net/

*Note: It would be highly recommended to use Google's reverse geocoding tool to get this information but unfortunately since it is now a paid service, we will have to use the tools at our disposal.*

4. We then use the **Folium package** to plot the zones on the map

5. Next, we use **Foursquare API** to get data for the popular spots in our zones

6. We use the **KMeans** module from **sklearn package** to cluster the data

*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.

The *K*-means clustering algorithm is used to find groups which have not been explicitly labeled in the data. This can be used to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets. Once the algorithm has been run and the groups are defined, any new data can be easily assigned to the correct group.

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.

Advantages:
K-means: Much faster than DBSCAN

DBSCAN: No number of clusters needed

Disadvantages:
K-means: Estimated of number of clusters required
DBSCAN: Does not work well over clusters with different densities

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

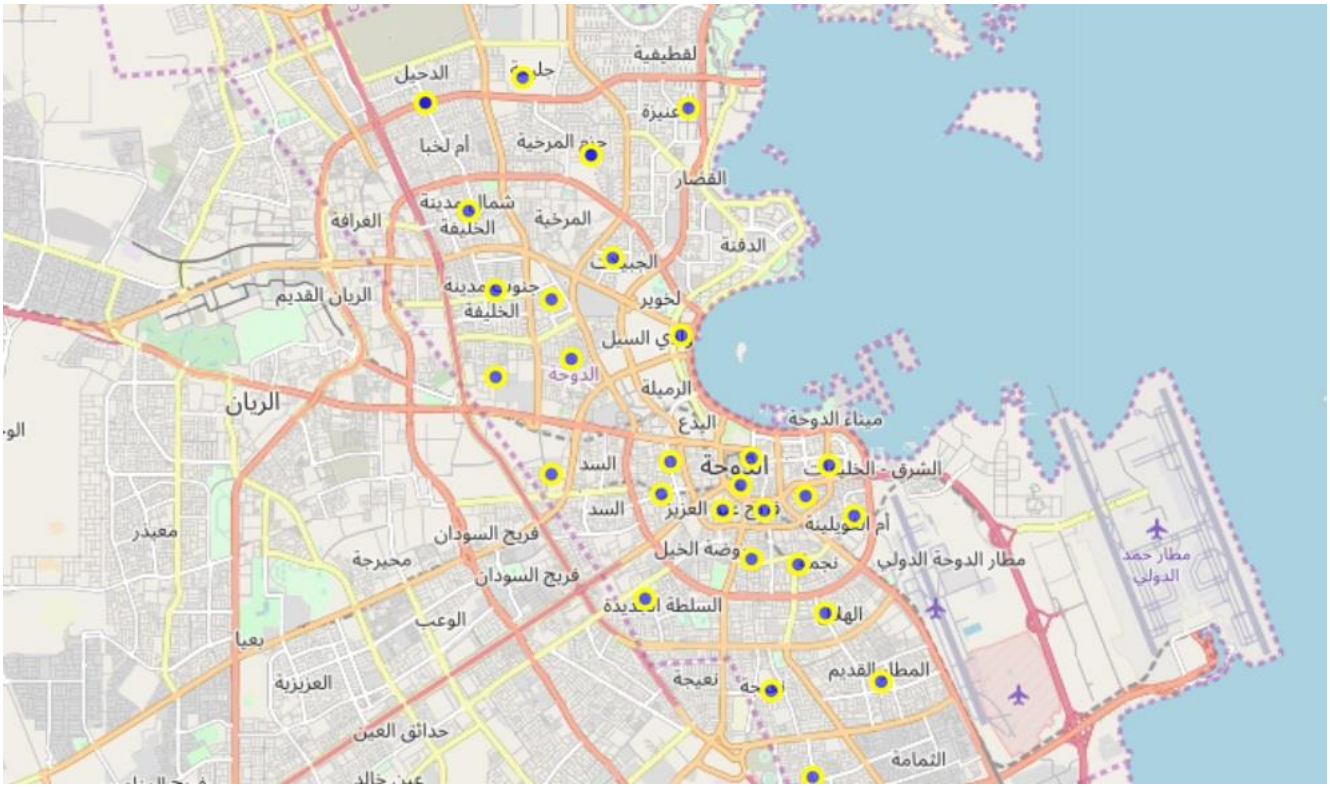
**Results**

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 |

We then visualize these areas on the map:




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

In [172]:
```
print(Venues.shape)
Venues
```

```
(900, 7)
```

Out[172]:

| | Neighborhood | Neighborhood Latitude | Neighborhood Longitude |
|---|---|---|---|
| 0 | Fereej Mohamed Bin Jasim | 25.2865 | 51.5296 |
| 1 | Fereej Mohamed Bin Jasim | 25.2865 | 51.5296 |
| 2 | Fereej Mohamed Bin Jasim | 25.2865 | 51.5296 |
| 3 | Fereej Mohamed Bin Jasim | 25.2865 | 51.5296 |
| 4 | Fereej Mohamed Bin Jasim | 25.2865 | 51.5296 |
| 5 | Fereej Mohamed Bin Jasim | 25.2865 | 51.5296 |
| 6 | Fereej Mohamed Bin Jasim | 25.2865 | 51.5296 |
| 7 | Fereej Mohamed Bin Jasim | 25.2865 | 51.5296 |
| 8 | Fereej Mohamed Bin Jasim | 25.2865 | 51.5296 |

With about 114 unique categories:

```
print('There are {} uniques categories.'.format(len(Venues['Venue Category'].unique())))
```

```
There are 114 uniques categories.
```

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=Fal
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
```

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

| | Zone | Districts | Population | Latitude | Longitude | Cluster Labels | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|
| **6** | 22 | Fereej Bin Mahmoud | 28327 | 25.2803 | 51.5124 | 0 | Café | Italian Restaurant | Coffee Shop |
| **9** | 26 | Najma | 28228 | 25.2683 | 51.5387 | 0 | Café | Hotel | Turkish Restaurant |
| **11** | 30 | Duhail | 7705 | 25.3477 | 51.4675 | 0 | Coffee Shop | Café | Shopping Mall |
| **22** | 42 | Al Hilal | 11671 | 25.2599 | 51.5439 | 0 | Café | Coffee Shop | Hotel |
| **26** | 63 | Onaiza | 37461 | 25.3469 | 51.5176 | 0 | Café | Beach | Italian Restaurant |
| **27** | 64 | Lejbailat | 4151 | 25.3212 | 51.5032 | 0 | Café | Coffee Shop | Indian Restaurant |

| | Zone | Districts | Population | Latitude | Longitude | Cluster Labels | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|
| **8** | 25 | Fereej Bin Durham | 37082 | 25.2693 | 51.5295 | 1 | Hotel | Turkish Restaurant | Café |
| **16** | 35 | Fereej Kulaib | 6507 | 25.3138 | 51.4914 | 1 | Café | Middle Eastern Restaurant | Spa |
| **18** | 37 | Fereej Bin Omran | 26121 | 25.3038 | 51.4953 | 1 | Restaurant | Coffee Shop | Hotel |
| **29** | 68 | Jelaiah | 5521 | 25.3522 | 51.4861 | 1 | Coffee Shop | Café | Supermarket |

| | Zone | Districts | Population | Latitude | Longitude | Cluster Labels | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 3 | Fereej Mohamed Bin Jasim | 4886 | 25.2865 | 51.5296 | 2 | Café | Hotel | Middle Eastern Restaurant |
| **5** | 17 | Al Rufaa | 6026 | 25.2853 | 51.5444 | 2 | Hotel | Restaurant | Middle Eastern Restaurant |
| **10** | 27 | Umm Ghuwailina | 33262 | 25.2766 | 51.5492 | 2 | Hotel | Restaurant | Café |
| **14** | 33 | Al Markhiya | 6242 | 25.3388 | 51.4992 | 2 | Café | Coffee Shop | Grocery Store |
| **20** | 40 | New Salatah | 16086 | 25.2623 | 51.5094 | 2 | Hotel | Coffee Shop | Department Store |
| **21** | 41 | Nuaija | 33379 | 25.2467 | 51.5334 | 2 | Coffee Shop | Café | Pizza Place |
| **25** | 61 | Al Dafna | 4022 | 25.3077 | 51.5163 | 2 | Café | Coffee Shop | American Restaurant |
| **28** | 67 | Hazm Al Markhiya | 8967 | 25.3388 | 51.4992 | 2 | Café | Coffee Shop | Grocery Store |

| | Zone | Districts | Population | Latitude | Longitude | Cluster Labels | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|
| **1** | 4 | Mushayrib | 28069 | 25.2818 | 51.5275 | 3 | Hotel | Mediterranean Restaurant | Middle Eastern Restaurant |
| **2** | 14 | Fereej Abdel Aziz | 15706 | 25.2777 | 51.5242 | 3 | Hotel | Pakistani Restaurant | Mediterranean Restaurant |
| **3** | 15 | Ad Dawhah al Jadidah | 15920 | 25.2776 | 51.5321 | 3 | Hotel | Café | Mediterranean Restaurant |
| **7** | 24 | Rawdat Al Khail | 18200 | 25.2860 | 51.5142 | 3 | Italian Restaurant | Café | Hotel |
| **15** | 34 | Madinat Khalifa South | 38247 | 25.3156 | 51.4808 | 3 | Café | Fast Food Restaurant | Middle Eastern Restaurant |
| **17** | 36 | Al Messila | 6803 | 25.3006 | 51.4808 | 3 | Café | Hotel | Restaurant |
| **23** | 45 | Old Airport | 48525 | 25.2481 | 51.5544 | 3 | Café | Coffee Shop | Fast Food Restaurant |

**Discussion**

Upon inspection of the clusters, its 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

Cluster #1

The first cluster is densely populated and has popular fitness and athletic spots. It also has a park which serves our purpose. The disadvantage is that this cluster also contains lots of cafes and shops that serve snacks.

Cluster #3

This is the smaller of the three clusters, however it does boast a park, shopping mall and gym amongst the popular spots. We can further investigate the locations of nearby schools in this area before making a decision.

Cluster #4

The fourth cluster is the most populated of all the clusters so that guarantees heavy foot traffic but it does lack gyms and fitness centers, so we cannot be certain if our target audience will be present in this cluster.

**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 😊