# The Battle of Neighborhoods in Dubai

This report is the part of capstone project of [IBM Data Science Professional Certificate](https://www.coursera.org/professional-certificates/ibm-data-science). The main aim of this blog post is to utilise all the concept we’ve learned from this certification for solving a business problem where we can use the Foursquare location data. Let’s see what are we going to solve.

## Table of Content

1. **Business Problem**
2. **Target Audience**
3. **Data Description**
4. **Methodology**
5. **Results & Outcomes**
6. **Conclusion**
7. **Business Problem:**

Expo 2020 Dubai is a World Expo that’s going to be hosted by Dubai in the United Arab Emirates. A World Expo is a mega international event in terms of size, scale, and duration and visitor numbers. It’s a festival and a platform where people from all over the world come together and connect with each other, share ideas, learn and innovate. It’s also a place you can come and have fun. Expo 2020 was scheduled on 20 October 2020 – 10 April 2021. and due to the COVID-19, the new schedule is 1 October 2021 – 31 March 2022. The staging of the world fair and the preparations leading up to it are expected to result an injection of nearly $40 billion into the economy, and an increase in visitors of at least 25 million persons from in and out UAE

1. **Target Audience:**

The dataset of Dubai has been used to help the visitors, investors and the job seekers to find suitable places such as restaurants, hotels, apartments, GYM's and so on.

1. **Data Description:**

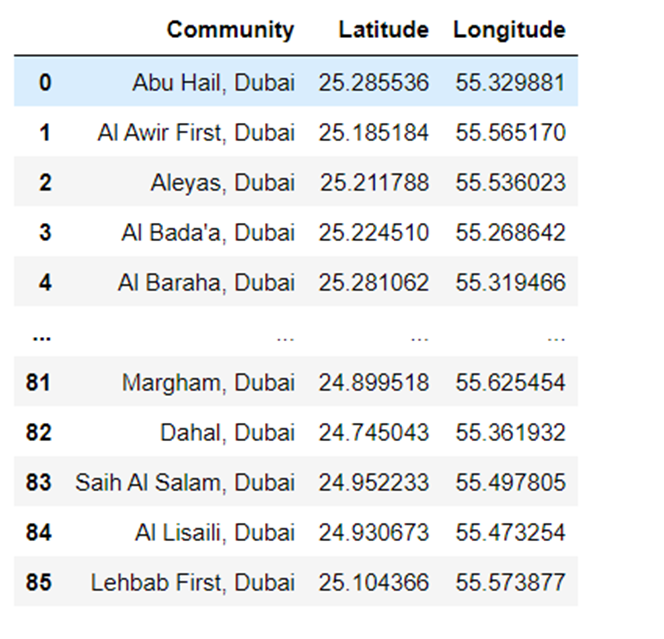
For this project we need the following data:  
*1. Dubai City data that contains Communities, Neighborhoods*

Data Source: <https://en.wikipedia.org/wiki/List_of_communities_in_Dubai>

Description: This Wikipedia page contain all the information we need to explore and cluster the neighborhoods in Dubai. We will be required to scrape the Wikipedia page and wrangle the data, clean it, and then read it into a *pandas* DataFrame so that it is in a structured format like the Dubai dataset.

*2. Geographical Location data using Geocoder Package*

Description: The second source of data provided us with the Geographical coordinates of the neighbourhoods.

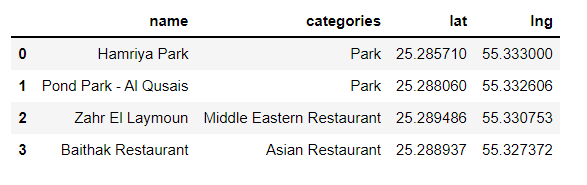


*3. Venue Data using Foursquare API*

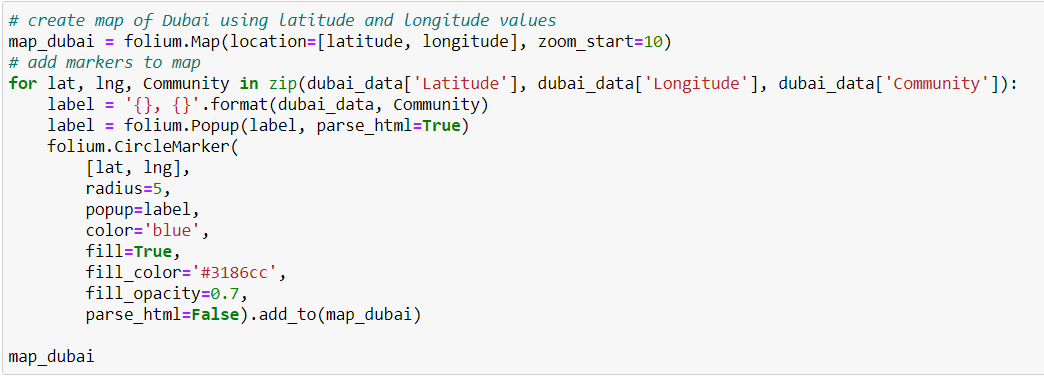
Data Source: <https://foursquare.com/developers/apps>

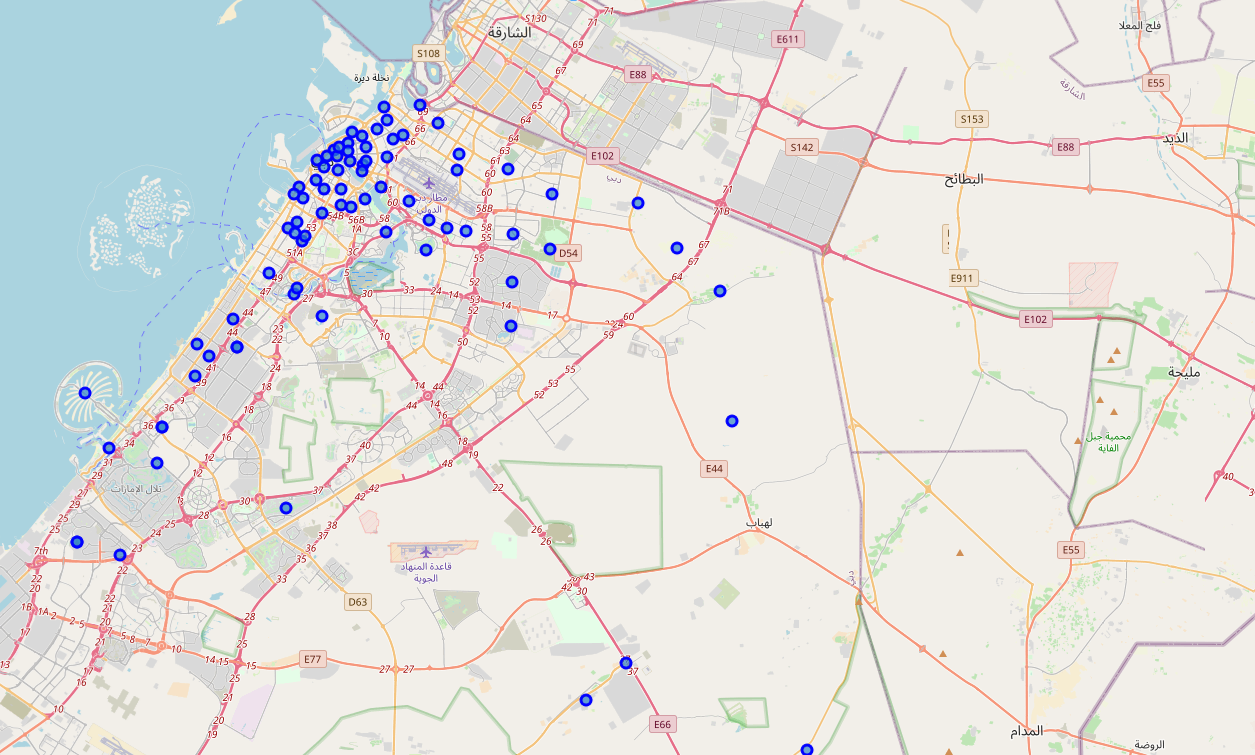
Description: From Foursquare API we can get the name, category, latitude, longitude for each venue.

1. **Methodology:**

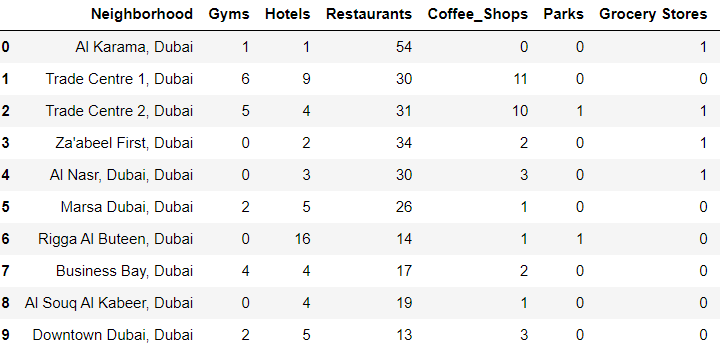
After scraping the data from Wikipedia there were Boroughs that were not assigned to any neighbourhood so we will get latitude and longitude of each borough and ignoring the boroughs whose data not available.

Now we will retrieve the venue data present within 500 meter radius of each neighborhood using Foursquare API and merge with the above table.

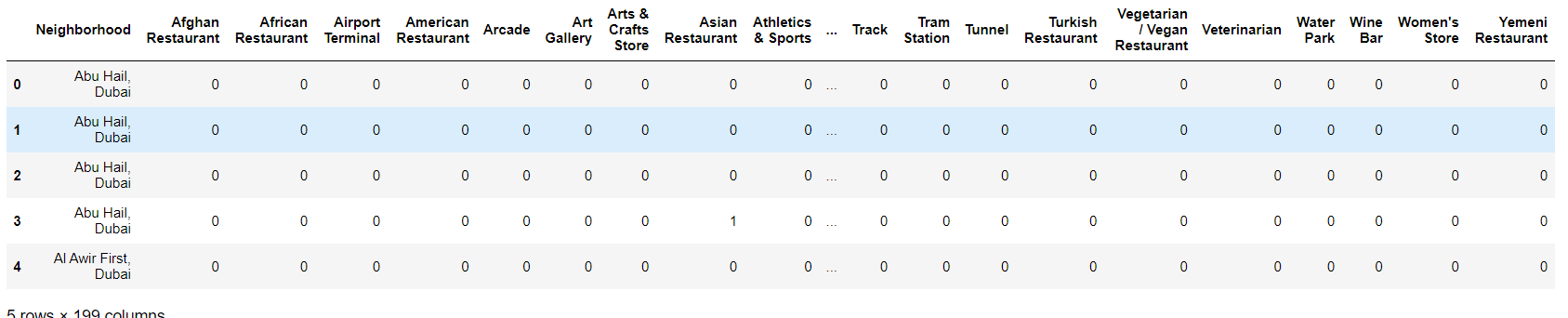
Now we need to visualise all neighborhoods in a map using Folium and color-coded each. The below bunch of code needed to do so.

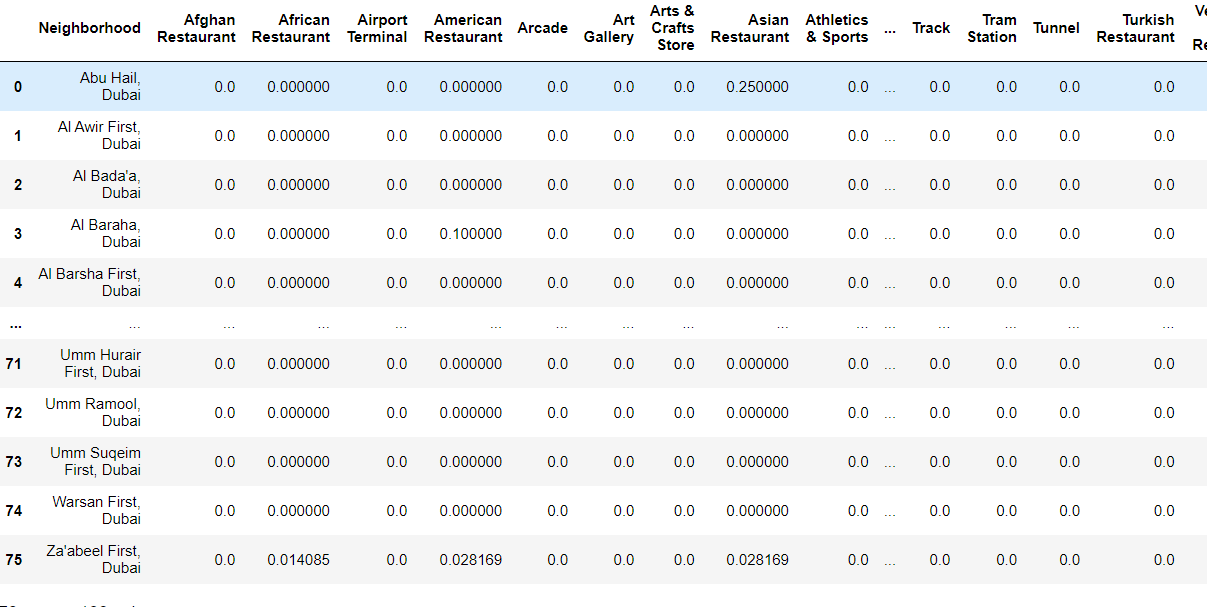
 This snippet of code provided us with the map below:

Next, we used the Foursquare API to get a list of all the Venues in Dubai which included Parks, Hotels, Coffee Shops, Restaurants etc. Getting this data was crucial to analyzing the number of all venues over Dubai Neighborhoods:



**Data Processing**

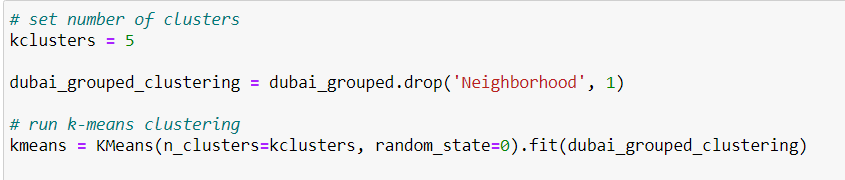
To analyze each neighborhood we'll use One hot encoding technique, individual venues were turned into the frequency at how many of those Venues were located in each neighbourhood.

Then we grouped those rows by Neighborhood and by taking the mean of the frequency of occurrence of each Venue Category.

After, we created a new DataFrame as well as the mean frequency of venues in that Neighborhood. This allowed the data to be summarized based on each individual Neighborhood and made the data much simpler to analyze.

**K-Means Clustering**

Now we’ll cluster these neighborhoods of Dubai based on the similarities of there venues using k-Means Algorithm (k-Means clustering was used to cluster the neighborhood into 5 clusters)

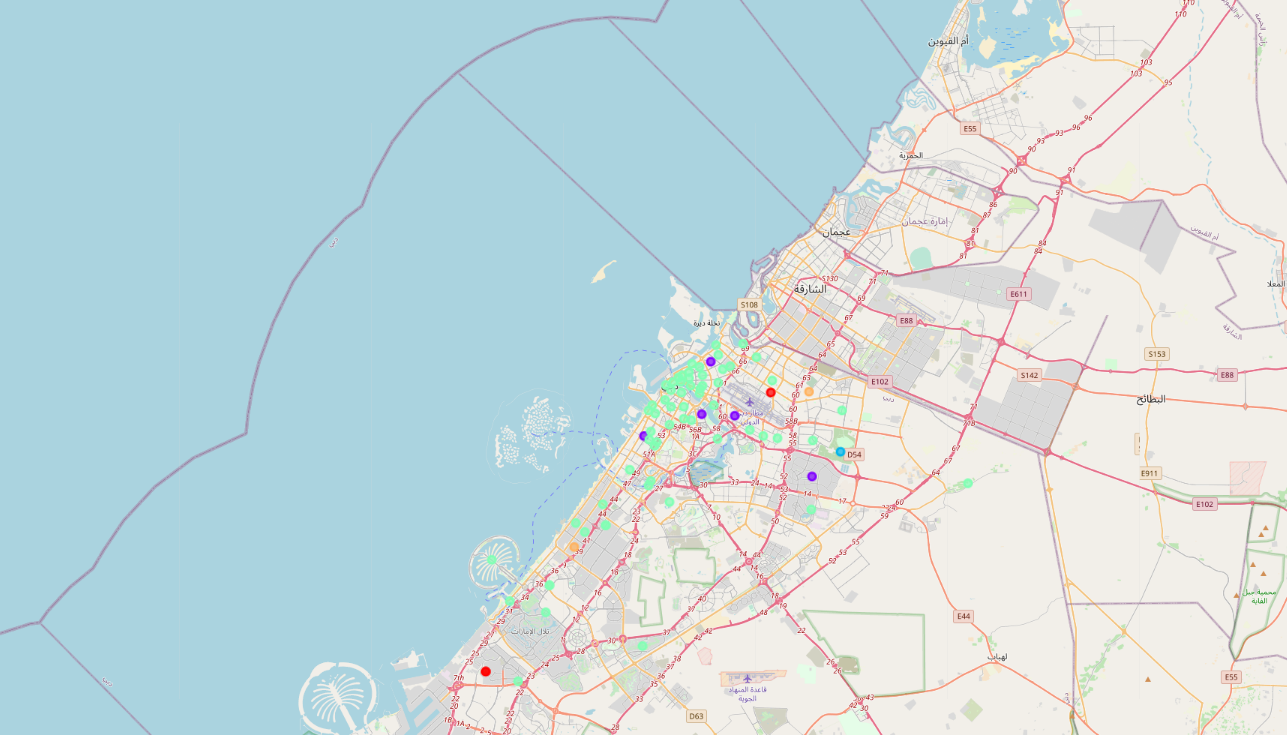


That means we will cluster the dataset into 5 cluster. Each of these clusters was labelled from 0 to 4 as the indexing of labels begins with 0 instead of 1.

1. **Result & Outcomes:**

Combining the results from the different analysis finally resulted the (cluster 1 & 0) shows an extended centre, with many restaurants, hotels, parks, and cafes.

The map below shows the different clusters that had a similar mean frequency of venues.



1. **Conclusion:**

This analysis gives a visitor a good indication of where to start his search for a suitable apartment. Although it doesn't guarantee that he will find the perfect place for his specific needs, it serves as a great supporting tool to narrow down his choices.

In conclusion, to end off this project, we had an opportunity on a business problem, and it was tackled in a way that it was similar to how a genuine data scientist would do.

To get the source code of this project click [here](https://github.com/alifarahat16/Capstone-Project/blob/master/The%20Battle%20of%20Neighborhoods%20in%20Duabi.ipynb).

Thanks for reading…