COURSERA CAPSTONE

IBM Applied Data Science Capstone

Opening a New coffee shop in Dubai,UAE

By: Alwaleed Alayyar

June 2020

Introduction:

This capstone project was developed to create a hypothetical scenario of an entrepreneur who wants to find the optimum location to open a coffee shop in Dubai city.

Business Problem:

The main goal of this project is to find the optimum location for the entrepreneur to open a new coffee shop in Dubai, UAE. By using data science methods and machine learning algorithims such as K-means clustering, this project aims to provide solutions to answer the business question: if an entrepreneur wants to open a coffee shop in Dubai, where should they consider opening it?

Target Audience:

This project is particularly useful to businessmen and investors looking to open or invest in new coffee shops in Dubai.

Data:

To solve the problem, we will need the following data:

• List of neighbourhoods in Dubai. This defines the scope of this project which is confined to the city of Dubai.

• Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and also to get the venue data.

• Venue data, particularly data related to coffee shops. We will use this data to perform clustering on the neighbourhoods.

Sources of data and methods to extract the data:

In this Wikipedia page (https://en.wikipedia.org/wiki/List_of_communities_in_Dubai) contains a list of neighborhoods in Dubai.

We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and

Beautifulsoup packages. Then we will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighborhoods.

After that, we will use Foursquare API to get the venue data for those neighborhoods. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the Shopping Mall category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used

Methodology:

Firstly, we need to get the list of neighborhoods in Dubai city. Fortunately, the list is available in Wikipedia.

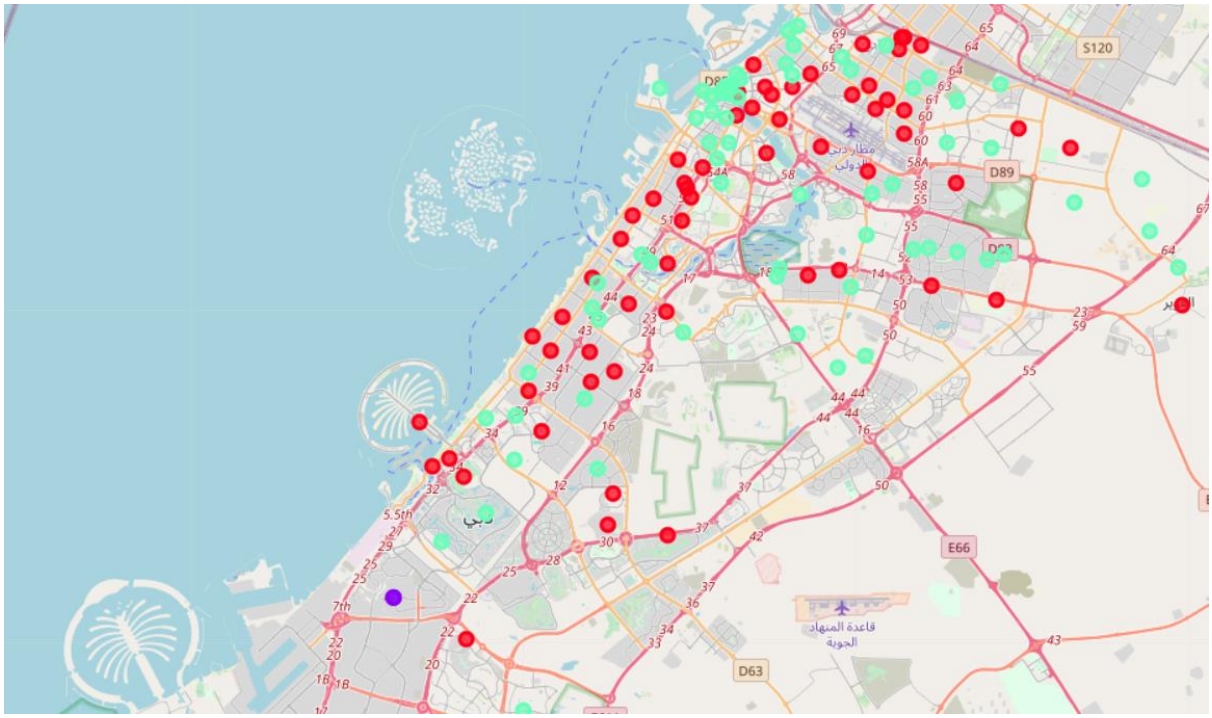
We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighborhoods data. However, this is just a list of the neighborhoods names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighborhoods in a map using Folium package.

This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in Dubai city.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 1000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analyzing the “coffee shop” data, we will filter the “coffee shop” as venue category for the neighborhoods.

Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighborhoods into 3 clusters based on their frequency of occurrence for “coffee shop”. The results will allow us to identify which neighborhoods have higher concentration of coffee shops while which neighborhoods have fewer number of coffee shops. Based on the occurrence of coffee shops in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new coffee shop.

Result:



The results from k-means clustering show that we can categorize Dubai neighborhoods into 3 clusters based on how many coffee shops are in each neighborhood:

- Cluster 1 (red): Neighborhoods with moderate number of coffee shops
- Cluster 2 (purple): Neighborhoods with high number of coffee shops
- Cluster 3 (green): Neighborhoods with low number to no existence of coffee shops