IBM Data Science Capstone Project

Battle of Neighborhoods (Week 1)
Dubai business nature

1. Description of the Problem and Discussion of the Background (Introduction Section):

One Foreign Investor is looking for opening a new investment project in Dubai as it is one of the most growing business cities in the area of the middle east,
But the problem is that our stockholders has a concept in his mind, but as being foreigner, he has not much idea about Dubai City structure and therefore needs help to determine what is the type of common businesses in Dubai the can do and what is the most rich and famous neighborhoods in Dubai with high population

Conclusion we have three main criteria's:

1- Which is common business fields in Dubai
2- What is the most rich and popular neighborhood
3- What are the target areas to implement the project?

Necessary Data and its usage in this case:

Part 1:

- List of all neighborhoods in Dubai.
- Total area of each neighborhood.
- Population of each neighborhood.
- Longitude and latitude of each neighborhood.

I have collected all of those data from Wikipedia and Dubai statistics center (DSC) web site and import it in csv file.

Part 2:

- List of venues in Dubai.
- Location of all venues (neighborhood).
- Type or the category of the business.
- Total number of each business category
- Total number of venues in each category.

As mentioned in IBM capstone course all of those data are available on foursquare API.

We will need data about different venues in different neighborhoods of that specific borough. In order to gain that information, we will use "Foursquare" locational information. Foursquare is a location data provider with information about all manner of venues and events within an area of interest. Such information includes venue names, locations, menus and even photos. As such, the foursquare location platform will be used as the sole data source since all the stated required information can be obtained through the API.

After finding the list of neighborhoods, we then connect to the Foursquare API to gather information about venues inside each and every neighborhood. For each neighborhood, we have chosen the radius to be 100 meters.

The data retrieved from Foursquare contained information of venues within a specified distance of the longitude and latitude of the postcodes. The information obtained per venue as follows.

Reference's:

https://www.dsc.gov.ae/en-us
https://developer.foursquare.com/docs/api-reference/venues/explore/
https://en.wikipedia.org/wiki/List_of_communities_in_Dubai

Methodology:

A Jupyter Notebook will be developed in order to process data and segment the neighborhoods. Following steps will be implemented:

1- Build neighborhoods list:

A list of districts is obtained from Madrid Wikipedia page. That list contains the names of the neighborhoods for every district.

2- Neighborhoods geolocation:

Every element in the neighborhood's dataset is geolocated using Python Geolocator and two columns are updated containing latitude and longitude for each district, neighborhood. The geolocator service has some problems, many times gives time out error, for this reason in this step the information obtained is saved in a text file in CSV format. Therefore, this step can be run many times, invoking geolocator only for missing data (timed out errors in previous executions). After various executions all the neighborhoods geolocation is obtained and we can use the text file.

As output a dataset containing a list of "district, neighborhood" is build.

3- Venues compilation:

As next step Foursquare services are used for obtaining venues for every neighborhood. The output is a new dataset with many records for every neighborhood containing the venues found for every one of them.
A free Foursquare service with limited count of calls is used. In order to minimize the usage of Foursquare, the information is saved in a text (CSV) file. It's supposed that the information gathered doesn't change in short period of time (some hours). When the analysis must be continued for long period (many hours or next day) just deleting the generated text file will force to call Foursquare services again, refreshing the information.

4- Neighborhoods segmentation:

The problem in hand is a case of unsupervised segmentation and, from the possible machine learning algorithms, K-means was chosen.
Taking in account that the venues information obtained from Foursquare is categorical, it must be previously processed in order to be handled by K-means algorithm. For this _"pandas.getdummies" is used for dummies variables. The list of dummy variables obtained are then grouped as features of every neighborhood. After executing K-means algorithm the "Elbow Curve" it's plotted in order to obtain the best K. Analyzing the change in the slope of the curve, it's determined that K=10 is a good value. K-means algorithm is executed. Next step is building the segmentation data frame, composed of the top venues for every neighborhood plus a segment label determined by K-means.

5- Segments analysis:

Every segment is printed individually, were different characteristics can be observed for each group. Next section describes the results

Results:

- Cluster 0 It's a big cluster where hotels and restaurant are the most common in the middle of the city
- Cluster 1 It's a small size cluster where the most common are entertainments activity like Zoo and golf
- Cluster 2 is middle size cluster with various categories like hotels, coffee shops and restaurants
- Cluster 3 This cluster is small cluster with most common with shopping malls, and various sport activity
- Cluster 4 This cluster is small cluster with various activity far from the city.

Most populated Areas:

1- Muhaisnah
2- Al Karama
3- Hor Al Anz
1- 4-Al Murraqabat

Most popular palces with Cafe/ Coffee shop:

1- Al Wasl
2- Al Khabisi
3- Umm Suqeim
2- 4-Jumeira

Most Common places with Restaurant:

1- Al Jaffiliya
2- Al Rigga/ Al Murraqabat
3- Abu Hail
4- Um Suqeim

Most Common places with Hotels:

1- Al Murar
2- Al Rigga
3- Al Baraha
4- Al Buteen

Discussion:

The objective of this project is found places in Dubai City for establishing the first investment project for a foreigner investor in the chain in such city. Main requirements are that other kind of restaurants exists and also entertainments for potential customers. Applying a given machine learning clustering algorithm was possible to segment neighborhoods based on their venues and, most important, found a group of them that have high potential. As a result, we are in condition of present our recommendations to the owner of the restaurant chain based on concrete data.

Conclusion:

We have gathered data from trusted sources and a known and strong methodology has been applied for processing A group of eleven neighborhood has been selected from more that on hundred that Dubai city has.

In such neighborhoods there are Hotels, restaurants and coffee shops. And also, Theaters and Soccer Fields can be found.