COURSERA

IBM Applied Data Science Capstone Project

Finding the best Location for Gas-Station in Dubai, UAE.

By

Nauman Mansuri

December, 2019

Introduction:-

With the booming automobile industry fuel consumption have tremendously increased to fulfil the need of commute. As a result provide great opportunity for oil company so build there gas station in Dubai city of UAE. In order to satisfy Endless fuel need. The company can earn a huge profit but it is not as easy at it seems location of Gas-station is to important before building. Not only location but there are several factors that affect the sales growth of a gas station Which needs to be taken to consideration.

Business Problem:-

The objective of this project is to select best location in Dubai, UAE to open a new Gas station using Data science technique and Machine learning algorithms like clustering which will be useful to provide solution for our Business problem.so if company wants to open their gas station in Dubai, UAE where you would advise to open?

Target Audience:-

This project is useful for the Oil companies and investors which are ready to invest in opening up a Gas-Station in the city of UAE, Dubai. This project is all time important for selecting the best geographical location according to Business need. As dubai has highly Concentrated Gas-Station in various areas and imbalance caused due to it will increase A chance to open up at a target location in order to hold strong advantage of it. According to report of gulf news Musaffah a area in UAE hits shortage of gas-station. Because of highly concentration to specific areas.

Data:-

Solving the problem will need following data

1. List of neighbourhoods in Dubai.

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

3. Venue data, particularly data related to Gas-Station. We will use this data to perform Kmeans clustering on the neighbourhoods.

Source: (https://en.wikipedia.org/wiki/List_of_communities_in_Dubai)

contains a list of neighbourhoods in Dubai with a total of 116 neighbourhoods. 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 neighbourhoods using python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods.

we will use Foursquare API to get the venue data for those neighbourhoods.. Foursquare API will provide many categories of the venue data, we are particularly interested in the Gas-station 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 ,Machine learning , 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. we would analyse it and than select nest cluster and location to achieve our target.

Methodology:

Firstly, we need to get the list of neighbourhoods in the city of Dubai. list is available in the Wikipedia page (https://en.wikipedia.org/wiki/List_of_communities_in_Dubai) .We will do web scraping using Python requests and Beautiful soup packages to extract the list of neighbourhoods data. However, this is just a list of 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 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 Data Frame and then visualize the neighbourhoods in a map using Folium package. This allows us to perform a check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of DUBAI.

we will use Foursquare API to get the top 100 venues that are within a radius of 2000 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 neighbourhoods 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 neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood 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 analysing the "Gas-Station" data, we will filter the "Gas-Station" as venue category for the neighbourhoods.

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 neighbourhoods into 3 clusters based on their frequency of occurrence for "Gas-Station". The results will allow us to identify which neighbourhoods have higher concentration of gas station while which neighbourhoods have fewer number of station. Based on the occurrence of station in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open gas_station.

Result

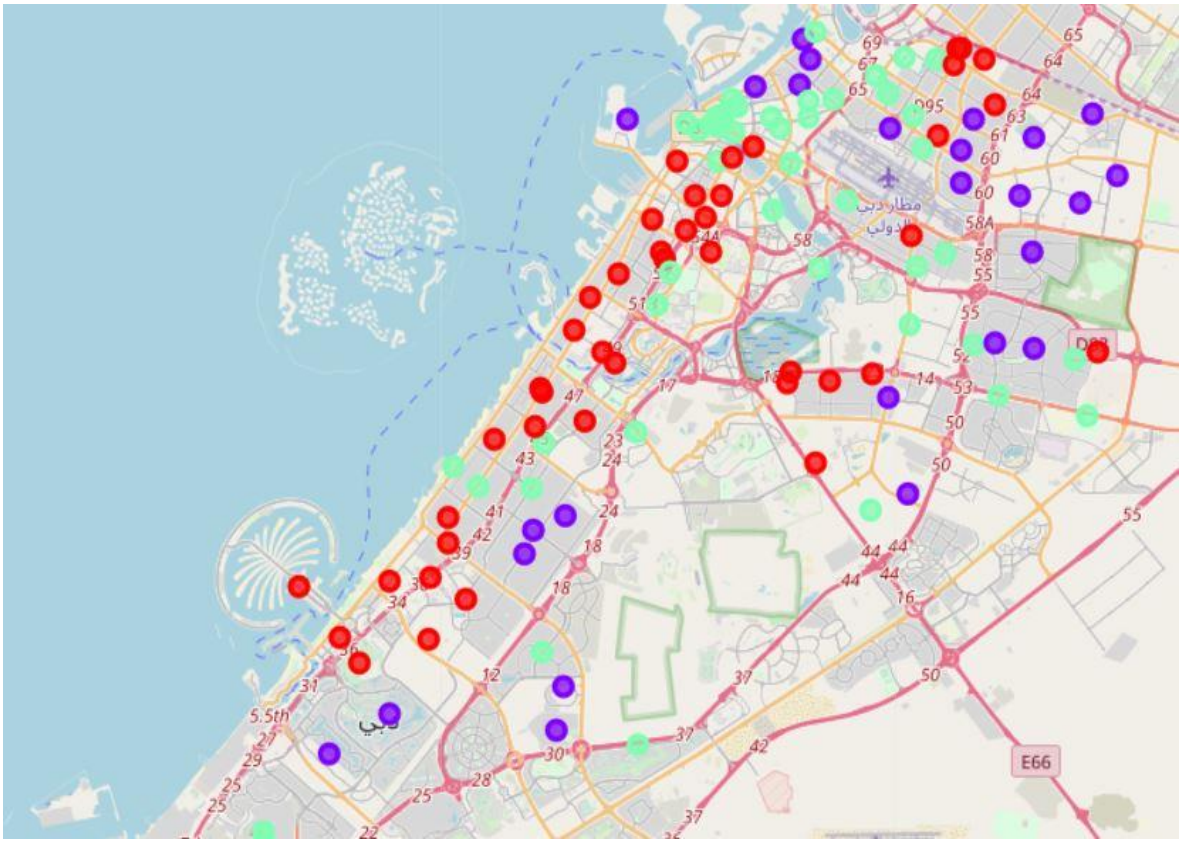
Below image shows the categorization of three clusters base on their frequency.

-Cluster 0 neighbourhoods with minimum number of gas station.
-Cluster 1 neighbourhoods with high number of gas station.
-Cluster 2 neighbourhoods with moderate number of gas station.

Cluster 0:-RED

Cluster 1:- purple

Cluster 2:-SEAGREEN



Discussion:

Form the observation obtained from map in the Results section, most of the stations are concentrated near industrial area of Kuala Lumpur city, with the highest number in cluster 1 and moderate number in cluster 2. On the other hand, cluster 0 has very low number to no Gas station
in the neighbourhoods. This represents a great opportunity and high potential areas to open new stations as there is very little to no competition from existing .Meanwhile, Station in cluster 1 are likely suffering from intense competition due to oversupply and high concentration From another perspective Therefore, this project recommends property oil company to capitalize on these findings to open stations in neighbourhoods in cluster 0 with little to no competition. Investors with unique selling propositions to stand out from the competition can also open new shopping malls in neighbourhoods in cluster 2 with moderate competition. Lastly, they are advised to avoid neighbourhoods in cluster 1 which already have high concentration of shopping malls and suffering from intense competition.

Limitations :

In this project, we only consider one factor i.e. frequency of occurrence of stations, there are other factors such as population and main roads that could influence the location decision of a new station. However, to the best knowledge of this researcher such data are not available to the neighbourhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a gas station.

Conclusion:

This project gone through the process of identifying business problem then specifying and scraping specific data set and performing clustering algorithm to group the neighbourhoods according to their similarity of number of Gas-Stations.so cluster 0 contains least station cluster2 contains moderate amount of cluster and cluster 1 contains maximum number of station. So this help oil companies to get best location to open their gas station and avoiding overcrowded areas for better growth.