Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Bakery in Dubai, United Arab Emirates

By: *Muhammad Bilal, Jan 2020*.

Introduction

For many shoppers, visiting Bakerys is a great way to relax and enjoy themselves during weekends and holidays. They can do grocery shopping, dine at restaurants, shop at the various fashion outlets, watch movies and perform many more activities. Bakerys are like a one-stop destination for all types of shoppers. For retailers, the central location and the large crowd at the Bakerys provides a great distribution channel to market their products and services. Property developers are also taking advantage of this trend to build more Bakerys to cater to the demand. As a result, there are many Bakerys in the city of Dubai and many more are being built. Opening Bakerys allows property

developers to earn consistent rental income. Of course, as with any business decision, opening a new Bakery requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the Bakery is one of the most important decisions that will determine whether the mall will be a success or a failure.

Business Problem

The objective of this capstone project is to analyse and select the best locations in the city of Dubai, UAE to open a new Bakery. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In the city of Dubai, UAE, if a property developer is looking to open a new Bakery, where would you recommend that they open it?

Target Audience of this project

This project is particularly useful to property developers and investors looking to open or invest in new Bakery in the major city of UAE i.e. Dubai. This project is timely as the city is currently suffering from oversupply of Bakery. Data from the National Property Information Centre (NAPIC) released last year showed that an additional 15 per cent will be added to existing Bakery space, and the agency predicted that total occupancy may dip below 86 per cent. Thelocal newspaper The Dubai Mail also reported in March last year that the true occupancy rates in malls may be as low as 40 per cent in some areas, quoting a Financial Times (FT) article cataloguing the country continued obsession with building more shopping space despite chronic oversupply.

Data To solve the problem

we will need the following data:

• **List of neighborhoods in Dubai**:

This defines the scope of this project which is confined to the city of Dubai, the capital city of the country of UAE in South East Asia.

• **Latitude and longitude coordinates of those neighborhoods**:

This is required in order to plot the map and also to get the venue data.

•**Venue data, particularly data related to Bakerys:**

We will use this data to perform clustering on the neighborhoods.

Sources of data and methods to extract them

This Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur) contains a list of neighborhoods in Dubai, with a total of 70 neighborhoods. 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 Bakery 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 the city of Dubai. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur). 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 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 wonderful 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 the city of Dubai. Next, 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 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 analyse each neighbourhood 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 analysing the “Bakery” data, we will filter the “Bakery” 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 “Bakery”. The results will allow us to identify which neighborhoods have higher concentration of Bakerys while which neighborhoods have fewer number of Bakerys. Based on the occurrence of Bakerys in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new Bakerys.

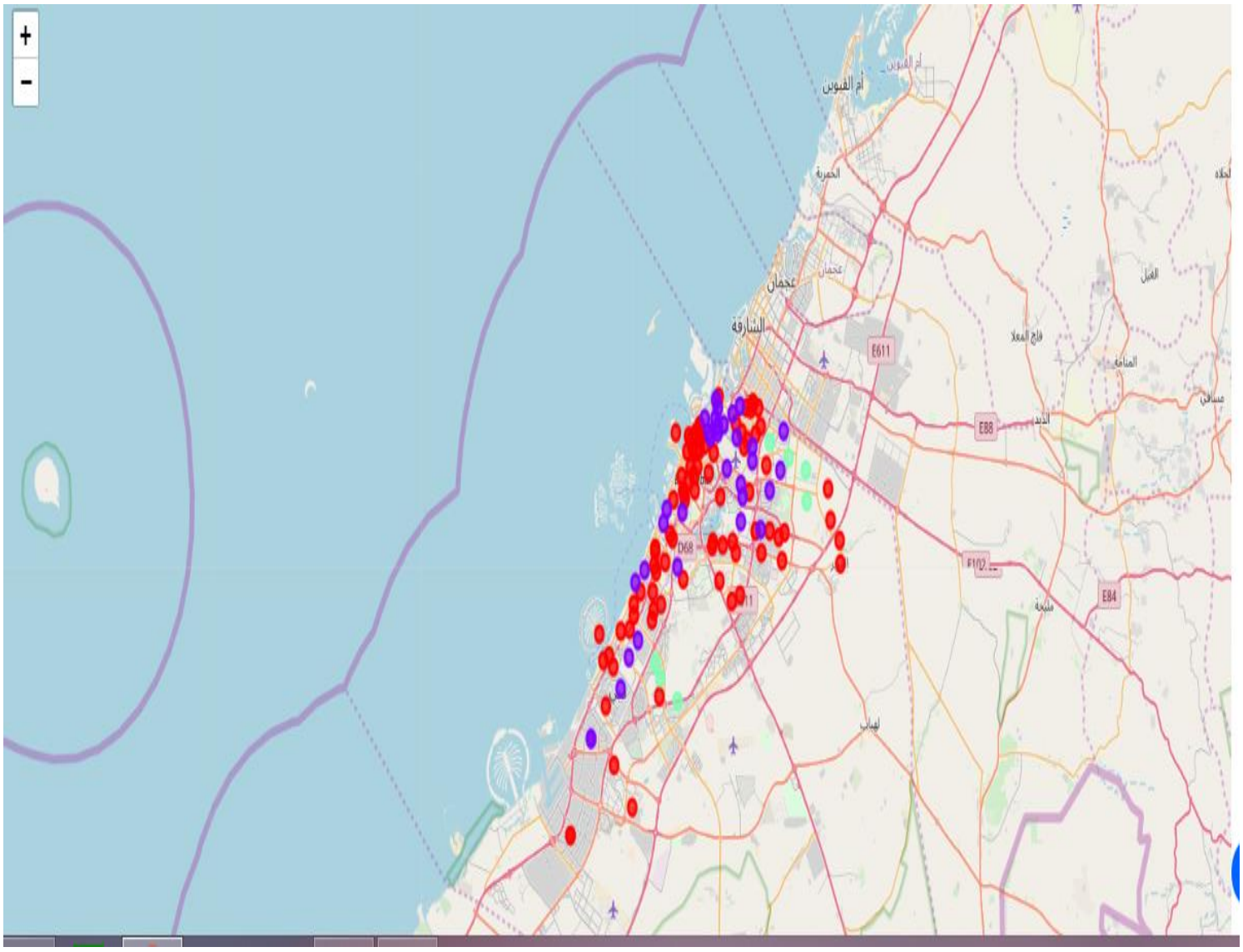
Results

The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for “Bakery”:

• Cluster 1: Neighborhoods with moderate number of Bakerys

• Cluster 2: Neighborhoods with low number to no existence of Bakerys

• Cluster 0: Neighborhoods with high concentration of Bakerys.



Discussion

As observations noted from the map in the Results section, most of the Bakerys are concentrated in the central area of Dubai city, with the highest number in cluster 2 and moderate number in cluster 0.

On the other hand, cluster 1 has very low number to no Bakery in the neighborhoods. This represents a great opportunity and high potential areas to open new Bakerys as there is very little to no competition from existing malls. Meanwhile, Bakerys in cluster 2 are likely suffering from intense competition due to oversupply and high concentration of Bakerys. From another perspective, the results also show that the oversupply of Bakerys mostly happened in the central area of the city, with the suburb area still have very few Bakerys. Therefore, this project recommends property developers to capitalize on these findings to open new Bakerys in neighborhoods in cluster 1 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new Bakerys in neighborhoods in cluster 0 with moderate competition. Lastly, property developers are advised to avoid neighborhoods in cluster 2 which already have high concentration of Bakerys and suffering from intense competition.

Limitations and Suggestions for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of Bakerys, there are other factors such as population and income of residents that could influence the location decision of a new Bakery. 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 new Bakery. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new Bakery. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 1 are the most preferred locations to open a new Bakery. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new Bakery.

References

Category:Suburbs in Dubai. Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Category:Suburbs_in_Dubai

Foursquare Developers Documentation. Foursquare. Retrieved from https://developer.foursquare.com/docs Malay Mail. (2018, March 14).

Malls facing meltdown as glut continues. Malay Mail. Retrieved from https://www.malaymail.com/s/1597735/malls-facing-meltdown-as-glut-continues Tan, H. H. (2018, April 5).

An oversupply of retail space in UAE. StarProperty.my. Retrieved from http://www.starproperty.my/index.php/articles/property-news/an-oversupply-of-retail-space-inUAE