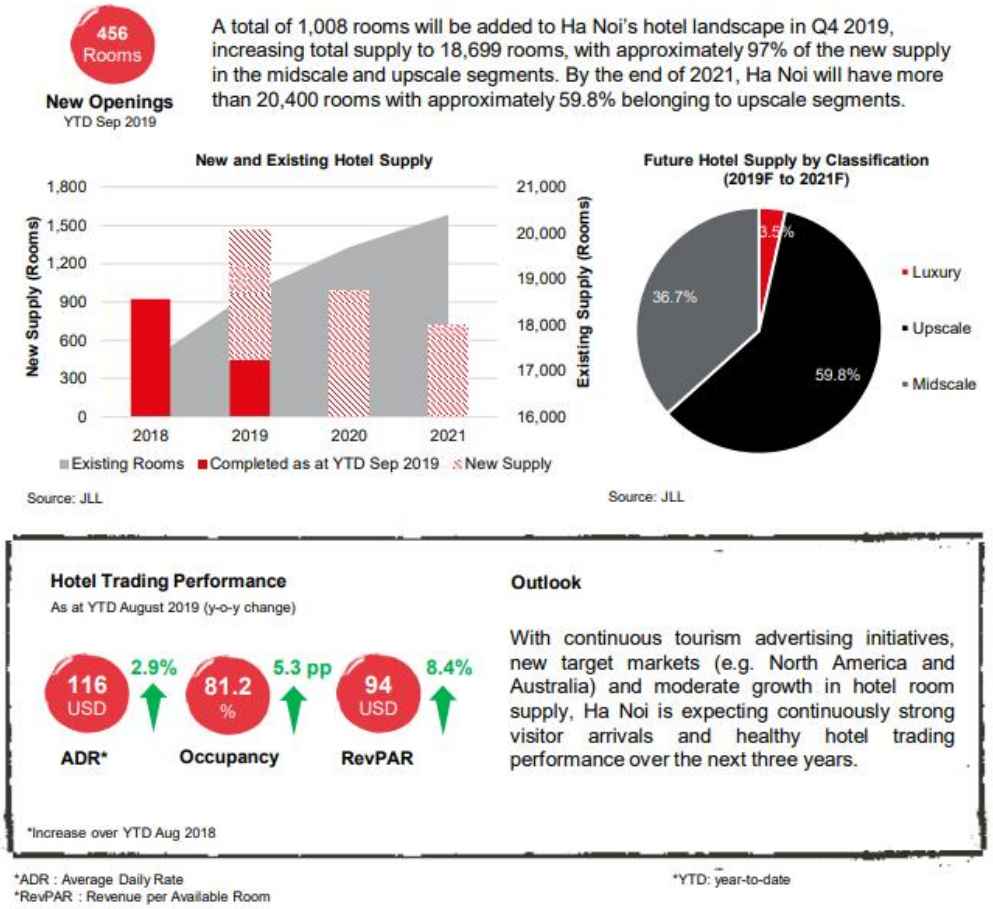
Coursera Capstone – Week 5 – Report

Opening a New Hotel in Ha Noi[1], Viet Nam

Problem

- The objective of this capstone project is to analyse and select the best locations in Ha Noi, Viet Nam to open a new hotel. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In Ha Noi, if a property developer is looking to open a new hotel, where would you recommend that they open it?
- This project is particularly useful to property developers and investors looking to open or invest in hotels in Ha Noi. This project is timely as the city is currently suffering from oversupply of hotels.
    - Data from Hotel news now[2]:
        - During the first two months of 2019, Hanoi reported double-digit increases in ADR (+10.1% to VND 2,856,659.12) and RevPAR (+14.0% to VND 2,200,763.03). Occupancy rose 3.5% to 77.0%, driven by a 4.1% jump in demand. The market experienced slight RevPAR growth in 2018 (+0.4%) once again pushed by ADR (+3.7%).
        - There are 224 hotels accounting for 17,615 rooms in Hanoi. The market continues to remain full of smaller hotels, with almost 75% of all hotels at 100 rooms or fewer and 60% of all hotels with fewer than 50 rooms. There are roughly 3,000 rooms in the development pipeline, with fewer than 1,000 rooms expected to open in 2019.
    - Data from JJL Vietnam[3]:



[1] https://en.wikipedia.org/wiki/Hanoi

[2] http://www.hotelnewsnow.com/Articles/294611/STR-Positive-outlook-for-Hanoi-hotel-industry

[3] https://www.joneslanglasalle.com.vn/content/dam/jll-com/documents/pdf/research/apac/vietnam/jll-vn-hotel-market-snapshot-hn-dn-hcmc-sept-2019-en.pdf

To solve the problem, we will need the following data

- List of neighbourhoods in Ha Noi. This defines the scope of this project which is confined to the city of Ha Noi.
- 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 hotels. We will use this data to perform clustering on the neighbourhoods.

Sources of data and methods to extract them

- This Wikipedia page (https://en.wikipedia.org/wiki/Category:Districts_of_Hanoi) contains a list of neighbourhoods in Ha Noi, with a total of 30 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.
- After that, 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 hotel 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).

Methodology

- Firstly, we need to get the list of neighbourhoods in Ha Noi. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/Category:Districts_of_Hanoi). We will do web scraping using Python requests and beautifulsoup 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 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 neighbourhoods 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 Ha Noi.
- Next, we will use Foursquare API to get the top 1.000 venues that are within a radius of 10.000 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 "Hotel" data, we will filter the "Hotel" 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 6 clusters based on their frequency of occurrence for "Hotel". The results will allow us to identify which neighbourhoods have higher concentration of hotels while which neighbourhoods have fewer number of hotels. Based on the occurrence of hotels in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open a new hotel.

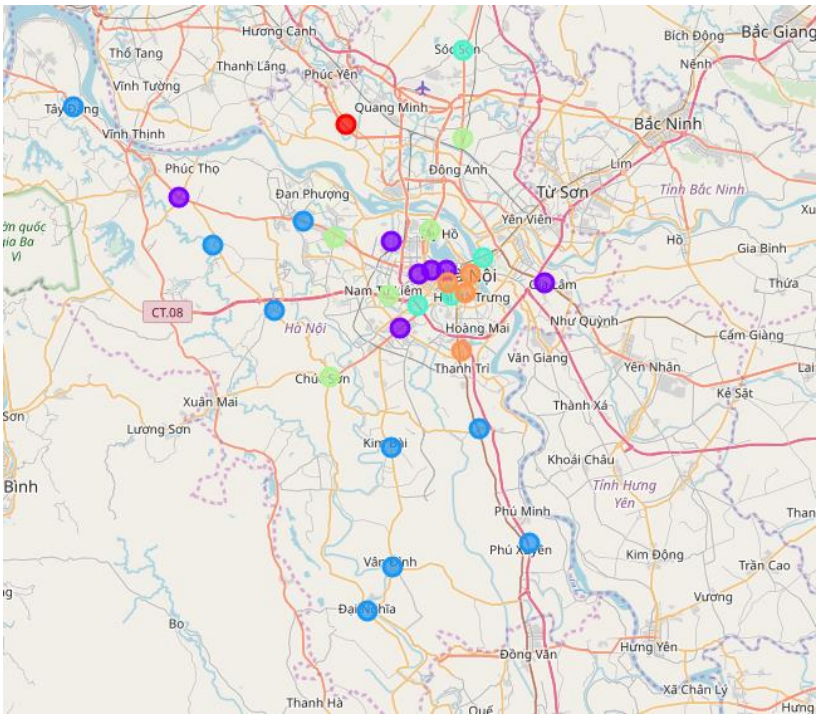
Results

The results from the k-means clustering show that we can categorize the neighbourhoods into 6 clusters based on the frequency of occurrence for "Hotel":

  - Cluster 2: Neighbourhoods with low number of hotel (blue)
  - Cluster 0: Neighbourhoods with low number of hotel (red)

- Cluster 1: Neighbourhoods with **medium** number of hotel (purple)
- Cluster 4: Neighbourhoods with **medium** number of hotel (green)
- Cluster 3: Neighbourhoods with **high** number of hotel (light blue)
- Cluster 5: Neighbourhoods with **high** number of hotel (orange)

	Neighborhood	Hotel	Cluster Labels	Latitude	Longitude
12	Mê Linh District	0.050000	0	21.182080	105.720610
1	Ba Đình District	0.150000	1	21.033520	105.814040
2	Bắc Từ Liêm District	0.150000	1	21.062170	105.769410
4	Cầu Giấy District	0.140000	1	21.029840	105.799530
5	Gia Lâm District	0.140000	1	21.019790	105.937510
19	Sơn Tây, Hanoi	0.150000	1	21.032796	105.830137
10	Hà Đông District	0.150000	1	20.973820	105.779160
16	Phúc Thọ District	0.142857	1	21.107110	105.537870
0	Ba Vì District	0.000000	2	21.199660	105.422700
26	Đan Phượng District	0.000000	2	21.083210	105.672810
24	Thạch Thất District	0.000000	2	21.058250	105.574950
23	Thường Tín District	0.000000	2	20.871610	105.865080
20	Thanh Oai District	0.000000	2	20.852820	105.768930
17	Quốc Oai District	0.000000	2	20.992210	105.641240
15	Phú Xuyên District	0.000000	2	20.754510	105.921020
29	Ứng Hòa District	0.000000	2	20.730550	105.771400
13	Mỹ Đức District	0.000000	2	20.685970	105.742760
11	Long Biên District	0.170000	3	21.045650	105.869640
18	Sóc Sơn District	0.166667	3	21.257320	105.848260
9	Hoàng Mai District, Hanoi	0.170000	3	21.007130	105.834910
22	Thanh Xuân District	0.170000	3	20.997740	105.798830
7	Hoài Đức District	0.125000	4	21.066100	105.707580
25	Tây Hồ District	0.110000	4	21.074180	105.812370
3	Chương Mỹ District	0.125000	4	20.923640	105.702680
27	Đông Anh District	0.125000	4	21.168130	105.848180
14	Nam Từ Liêm District	0.120000	4	21.008130	105.766500
8	Hoàn Kiếm District	0.160000	5	21.029020	105.856220
21	Thanh Trì District	0.160000	5	20.951230	105.846210
6	Hai Bà Trưng District	0.160000	5	21.009910	105.850760
28	Đống Đa District	0.160000	5	21.020410	105.830820



Discussion

As observations noted from the map in the Results section, most of the hotels are concentrated in the central area of Ha Noi, with the highest number in cluster 3&5 and moderate number in cluster 1&4. On the other hand, cluster 0&2 has very low number to no hotel in the neighbourhoods. This represents a great opportunity and high potential areas to open a new hotel as there is very little to no competition from existing malls. Meanwhile, hotels in cluster 3&5 are likely suffering from intense competition due to oversupply and high concentration of hotels. From another perspective, the results also show that the oversupply of hotels mostly happened in the central area of the city, with the suburb area still have very few hotels. Therefore, this project recommends property developers to capitalize on these findings to open a new hotel in neighbourhoods in cluster 1&4 with little to no competition (avoid neighbourhoods in cluster 3&5 and too far from center area in cluster 0&2).

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 6 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 hotel. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 1&4 are the most preferred locations to open a new hotel. 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 hotel.