# Coursera Capstone

IBM Data Science Professional Certificate

## *Opening a new hotel in Hanoi, Vietnam*

Duc Vu - Feb 2020

# Business Problem

Vietnam is world renowned for its natural beauty, rich history and amazing cuisine. As described by Lonelyplanet: *"A land of staggering natural beauty and cultural complexities, of dynamic megacities and hill-tribe villages, Vietnam is both exotic and compelling"*. With this in mind, it is no surprise that tourism is a booming sector in Vietnam, with millions of travellers from all over the world choosing the country as their holiday destination every year. As a result, the hotel industry of Vietnam is a very attractive field for investors. However, it is a very competitive field given the number of hotels already existing and the ones being built. This is especially true for Hanoi, the capital city of Vietnam.

As with any business decision, opening a new hotel requires serious planning beforehand, one of which is choosing the location for the venue. This is crucial in determining whether the hotel will be a successful or failed investment.

This capstone aims to analyse and select the best possible locations to open a new hotel in Hanoi, Vietnam using data science.

# Target Audience

This project is useful for current and potential developers or investors looking to step into the hotel market of Hanoi, Vietnam.

# Data

**List of neighbourhoods in Hanoi**: This can be scraped from Wikipedia. This data defines the scope of this project which is confined to the city of Hanoi, Vietnam. The aim for now is to explore the data on a Ward level, where English resources are not readily available thus the use of Vietnamese sources. If this doesn't work out during the project we can move one level up to the District level where more English data is available.

**Latitude and longitude coordinates of those neighbourhoods:** This is required for data to be plotted on a map. The data can be retrieved through Python's Geocode package .

**Venue data, especially those related to hotels**: Clustering will be performed based on this data. The data can be retrieved using Foursquare API.

# Methodology

Firstly, we need to get the list of districts in the city of Hanoi. Fortunately, the list is available in this Wikipedia page. We will do web scraping using Python requests and the beautifulsoup package. 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 use Foursquare API. This can be achieved using the Geocoder package. After gathering the data, we will populate the data into a pandas DataFrame then visualize the districts on a map using the 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 Hanoi.

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 districts 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 district and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each district by grouping the rows by district 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”, “Hostel” and “Motel” as venue categories for the districts. We will also drop districts with 0 frequency (too rural).

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 districts into 3 clusters based on their frequency of occurrence for the different hotel categories. The results will allow us to identify which districts have higher concentrations of hotels while which districts have fewer number of hotels. Based on the occurrence of hotels in different districts, it will help us to answer the question as to which districts are most suitable to open new hotels.

# Result

The results from the k-means clustering show that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for “Hotel”, “Hostel” and “Motel”:

- Cluster 0: Districts with moderate number of hotels - Ba Dinh, Hoan Kiem, Me Linh
- Cluster 1: Districts with high concentration of hotels - Cau Giay, Hai Ba Trung, Nam Tu Liem, Tay Ho, Dong Da
- Cluster 2: Districts with close to non-existence of hotels - Hoang Mai

# Discussion

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

# 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. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The districts in cluster 2 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 new hotel.