IBM APPLIED DATA SCIENCE CAPSTONE PROJECT

Recommend the best location to start a business in Hanoi

ABSTRACT

Hanoi, a thousand-year-old capital city of Vietnam, is well-known for its historical heritage and beautiful landscape. Thanks to that, investors have found Hanoi as an attractive destination, thus, more and more businesses are opened in this city. Considering this trend, the aim of the project is to sort out the best location to start a business in Hanoi.

LINH THUY PHUNG

June, 2020

Introduction

As one of the most ancient cities in the world that is filled with rich culture and beautiful lakes, Hanoi is the capital of Vietnam that is situated in the northern part of the country. For more than a thousand year, Hanoi has always been the center of culture and political, creating a deep connection with government and large network for investors to share information and experiences. Moreover, the city is chosen as headquarter of many educational institutions, combined with a high proportion of youngster, it naturally provides a large number of high-quality young labors. Especially, the inexpensive labor costs is particularly advantageous to small businesses in saving costs and boosting revenues. In addition to that, Vietnamese government has made a lot of efforts on the openness of foreign investment, reformed policies, welcoming business environment, and simplified process to lower the barriers for both domestic and foreign investors. For these reasons, it is no doubt that Hanoi has become the hot spot for investment and starting new business.

Business problem

Starting a new business is never easy as it requires a thorough plan with so many factors to be taken into consideration. Among them, choosing a location is one of the most important decision. It especially arduous for foreigners who have little knowledge of the local destination. Therefore, the objective of this capstone project is to solve that specific task. In this project, we shall analyze data to select the most suitable location in Hanoi to open a business. Using data science knowledge, the project aims to explore geographical data of Hanoi, make visualization with map chart and marker and then, cluster venues into groups to find out which group is the best to locate a new business.

Target audience of the project

As previously mentioned, the main user of this project will be investors who are interested in opening a business in Hanoi, Vietnam. Among them, foreigners would be more keen on the results due to their unfamiliarity with the local background. Moreover, the results are not solely restricted to one particular viewer like investor, it could also yield benefits for many other subjects

Local authorities could use it to understand which part of the city is the most occupied with businesses to manage. Builders can seek the information of the most popular area to construct new building for renting. Even students can find it useful to find the location of potential employer.

Data section

Introduction of data to be used to solve the business problem

A list of all neighborhoods in Hanoi. The scope of this project therefore will be restricted to only Hanoi city of Vietnam, since it is the special location to attract investment and is currently home of many businesses

The coordinate data of Hanoi city, including specific latitude and longitude of all neighborhood in the above list. This information will eventually be used to plot a map chart to depict and visualize the location of each neighborhood

Data of venues around the neighborhoods in Hanoi. At the end of the project, these venues shall help us cluster the neighborhoods into different groups

Data source and methodology to retrieve data

Wikipedia is a multilingual online encyclopedia created and maintained as an open collaboration project by a community of volunteer editors using a wiki-based editing system. It is the largest and most popular general reference work on the World Wide Web. Hence, For the list of neighborhoods in Hanoi, we will use the data extracted from Wikipedia. The list contains 30 neighborhoods could be retrieved from this link: https://en.wikipedia.org/wiki/Category:Districts_of_Hanoi. Then, we shall scraping data from this website using BeautifulSoup.

After that, we install the Geocoder package of Python. With the names of all neighborhoods in Hanoi city, Geocoder could help identify the geographical coordinates, including both latitude and longitude.

Finally, Foursquare API could be used in this last step to gather all venues around the city. It has one of the largest databases of information about more than 105 million locations, thus, it becomes really famous and frequently used by many developers around the world. So, in this project, we also use Foursquare API to retrieve data of venues located in Hanoi. The detailed explanation on methodology will be discussed in the next section.

Methodology

Firstly, we must import all the necessary package and libraries as follows:

- Pandas: library for data analysis

- Requests: library to handle requests

- BeautifulSoup: library to parse HTML and XML documents

- Geocoders: package to convert an address into latitude and longitude values

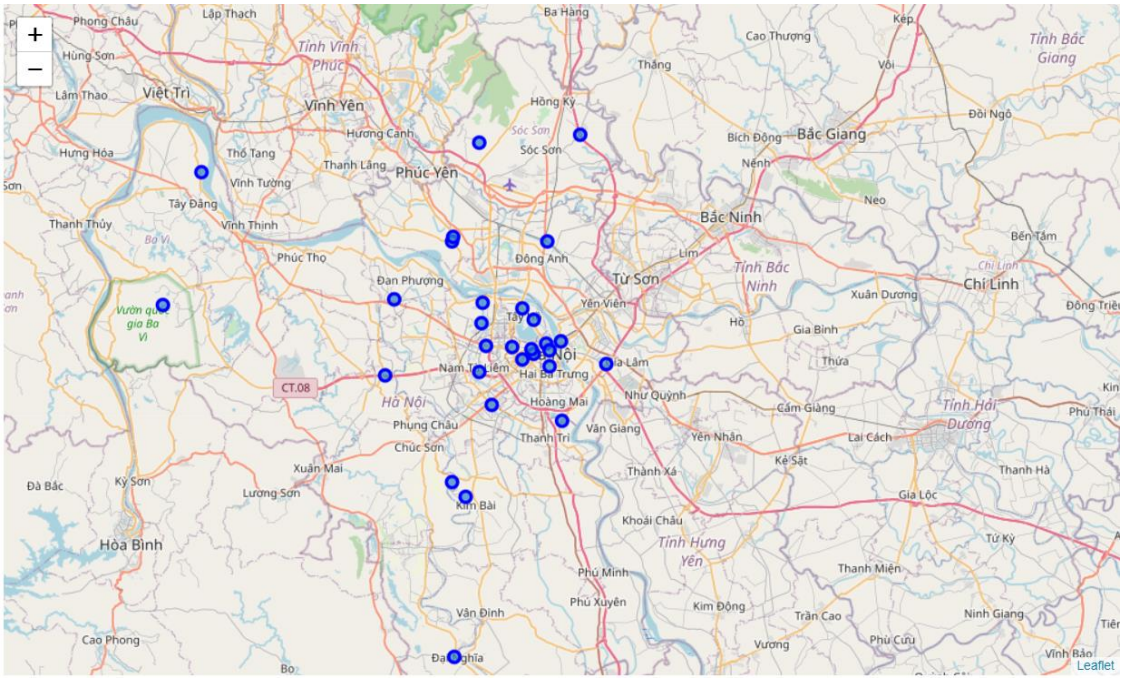
- Matplotlib: library to plot chart

- Folium: map rendering library

- Kmeans from sklearn: used for clustering venues

Then, we shall use the link: https://en.wikipedia.org/wiki/Category:Districts_of_Hanoi from Wikipedia to extract the list of neighborhoods in Hanoi with Requests and BeautifulSoup. After that, a new dataframe will be created from the list using pandas library. The purpose of this action is to serve the cleaning and analyzing data later.

The next step is getting geographical coordinates for each neighborhood. Using the name of neighborhoods in the city as the basis, we can pull off their latitude and longitude with Geocoder package. Then, add them as new columns to the previous Dataframe. Now, we manage to get the name of all neighborhoods in Hanoi, along with their coordinates. From this, plot a map of Hanoi city with each neighborhood is presented as a marker to check whether the extracted information is correct. Also, a visualization would help us to quickly understand the data better.



At first glance, we can easily spot all neighborhoods in the above map. It is clear that most of them are distributed around the center, only a few is located far away. In other words, there is a possibility that this area could be the ideal spot to attract more people. However, this is just an assumption so we must examine further.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 500 meters. Each venue will be populated with following information: Venue name, Venue Latitude, Venue Longitude and Venue Category. Then, we convert them into a new dataframe, check how many venues returned for each Neighborhood and find out how many unique categories can be curated from all the returned venues by using pandas. With this data, we shall analyze the neighborhoods by group rows by neighborhood and by taking the mean of the frequency of occurrence of each category. However, with too many returned venues, it is difficult to analyze. Therefore, to simplify, we shall filter our data to get only top 10 venues for each neighborhood.

Finally, based on the above cleaned data, we will run k-means to cluster Hanoi into 4 clusters. The k-means algorithm searches for a pre-determined number of clusters within an unlabeled multidimensional dataset. It accomplishes this using a simple conception: The "cluster center" is the arithmetic mean of all the points belonging to the cluster and each point is closer to its own cluster center than to other cluster centers. With this feature, the k-means is particularly suitable for solving the business problem. The results will allow us to identify which neighborhoods have higher concentration of businesses, and from there, it helps us answer the question that which area is the best location to open a business in Hanoi.

Results

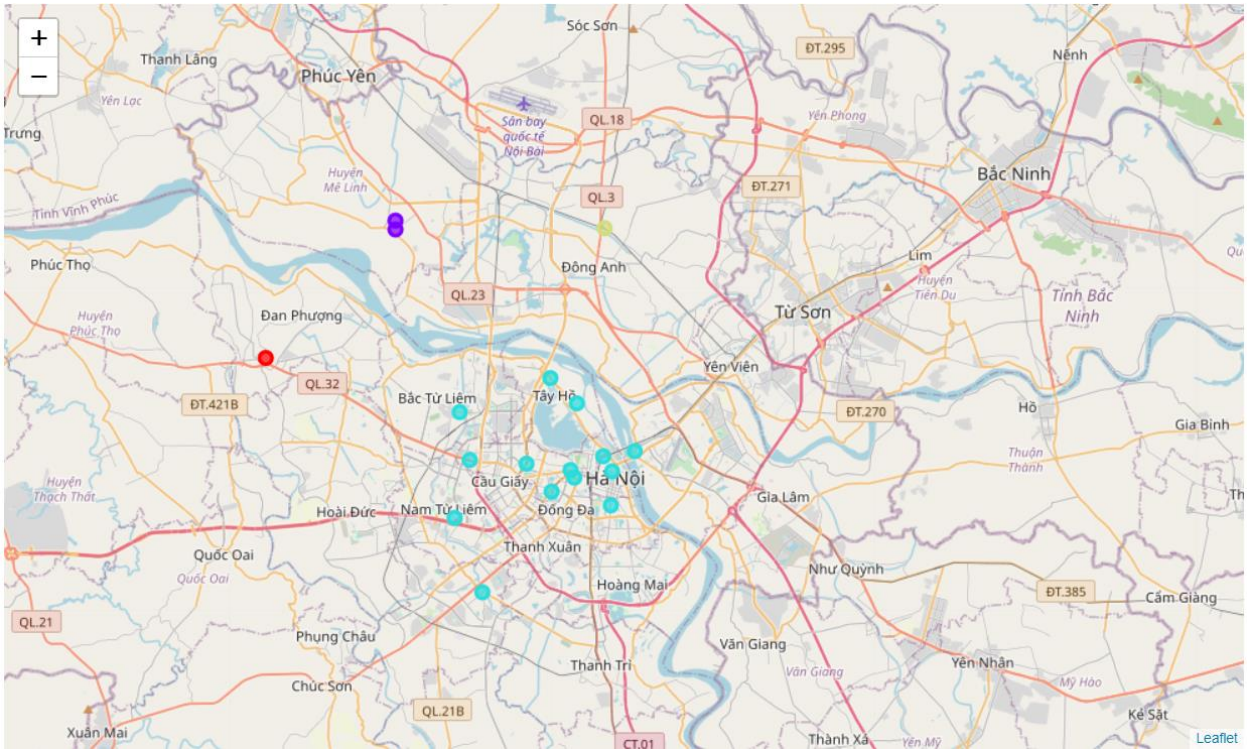
After running k-means to our data the results show that we classify neighborhoods into 4 clusters based on the frequency of occurrence as follows:

- Cluster 0 (red): Neighborhoods with small to no existence of businesses

- Cluster 1 (violet): Neighborhoods with moderate number of businesses

- Cluster 2 (blue): Neighborhoods with large number of businesses

- Cluster 3 (yellow): Neighborhoods with small to no existence of businesses



Summary

Discussion

From the results presented above, we can confidently confirm the theory that the center will be the most suitable location to open a new business. This statement is proven with the fact that cluster 2 which are located around the center has the highest frequency of occurrence of businesses. Normally, people want to avoid areas which are already populated with stores because of the high competition. However, it is not the case of Hanoi, since high number of businesses doesn't mean we should avoid that, instead, investors should do more research to find the unique business niche. Hanoi is a rich cultural capital city, so traditional jobs has been here for a long time. It would be a great opportunity for new type of businesses show up. If we look back into the history of establishment of Hanoi city, this suggestion is strengthened even more. In the past, Hanoi area only include those neighborhoods in the center. The surroundings which are within cluster 0, 1 and 3 are used to be agricultural land and now, part of them become factories. Hence, if you want to open a business in Hanoi, it is recommended to open office/headquarter in area of cluster 2 and place the factories in the remaining clusters.

Conclusion

As observed from the map, cluster 2 has the largest number of businesses, while cluster 0 and cluster 3 have small to no existence of business, and cluster 1 only have a few businesses. Because of a long history of traditional businesses in Hanoi, we suggest that cluster 2, though seems to be crowded, to be the best location to open business if investors seek for a dynamic area to try something new and unique. At the same time, the remaining areas also should be exploited, but for different purpose, to place factories or plant natural resources. In the end, the project only gives us an overview of the area to suggest the most suitable location, but neglecting many other factors, such as population and average income. Therefore, future research based on this result should be conducted with more factors involved to overcome limitations of this project.