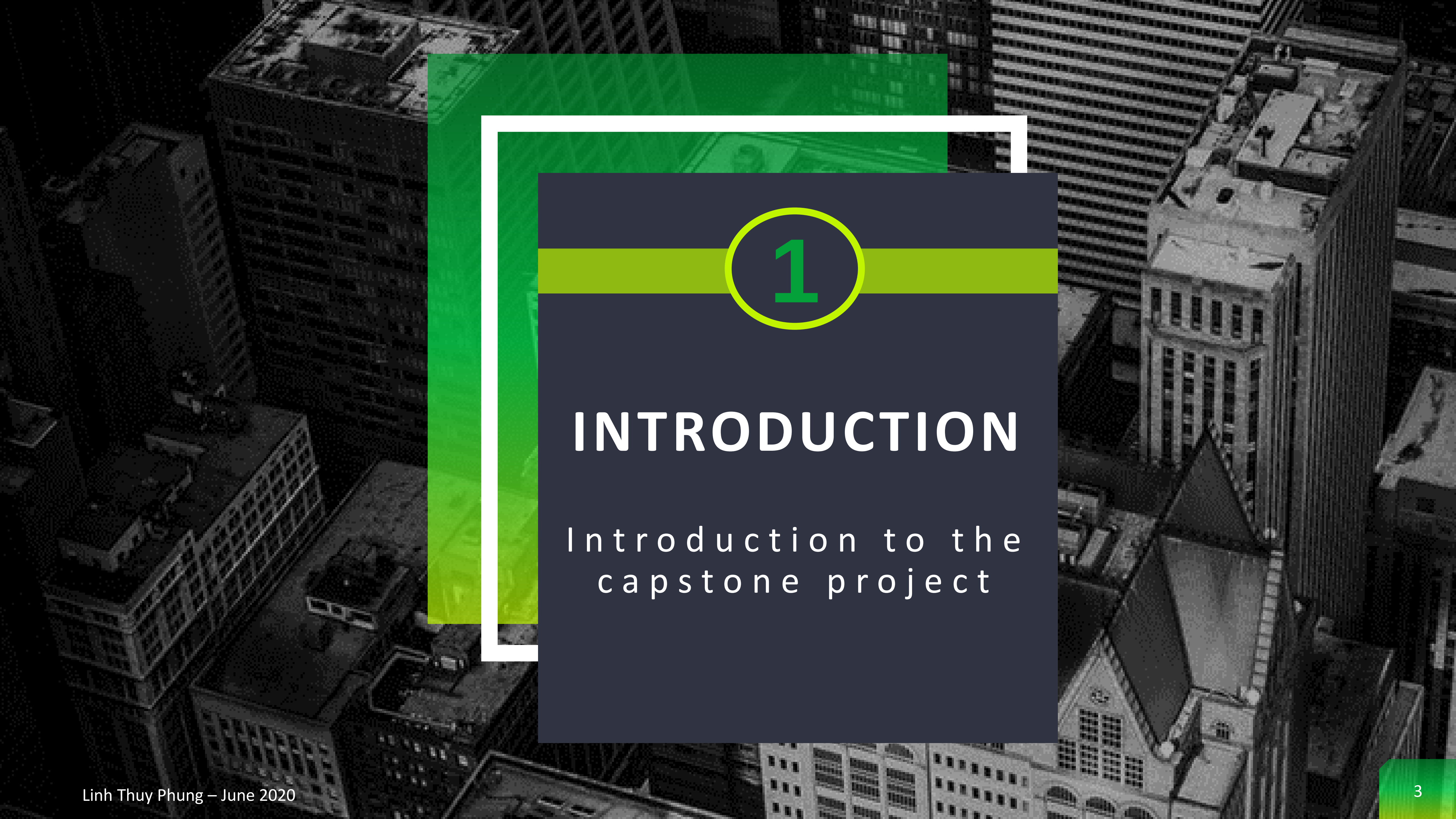
IBM APPLIED DATA SCIENCE CAPSTONE PROJECT

Recommend the best location to start a business in Hanoi

By Linh Thuy Phung

CONTENTS

1. Introduction
   - Business problem
   - Target audience
   - Data
2. Methodology
3. Results
4. Summary
   - Discussion
   - Conclusion

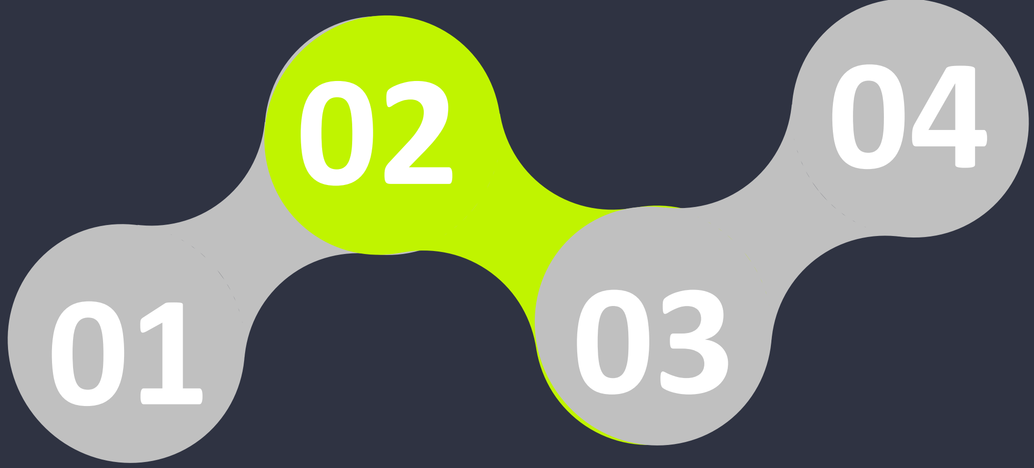


1

INTRODUCTION

Introduction to the capstone project

WHY CHOOSE HANOI ?



01
Center of culture
and political

02
Large
networks

03
High-quality
young labors

04
Government
support

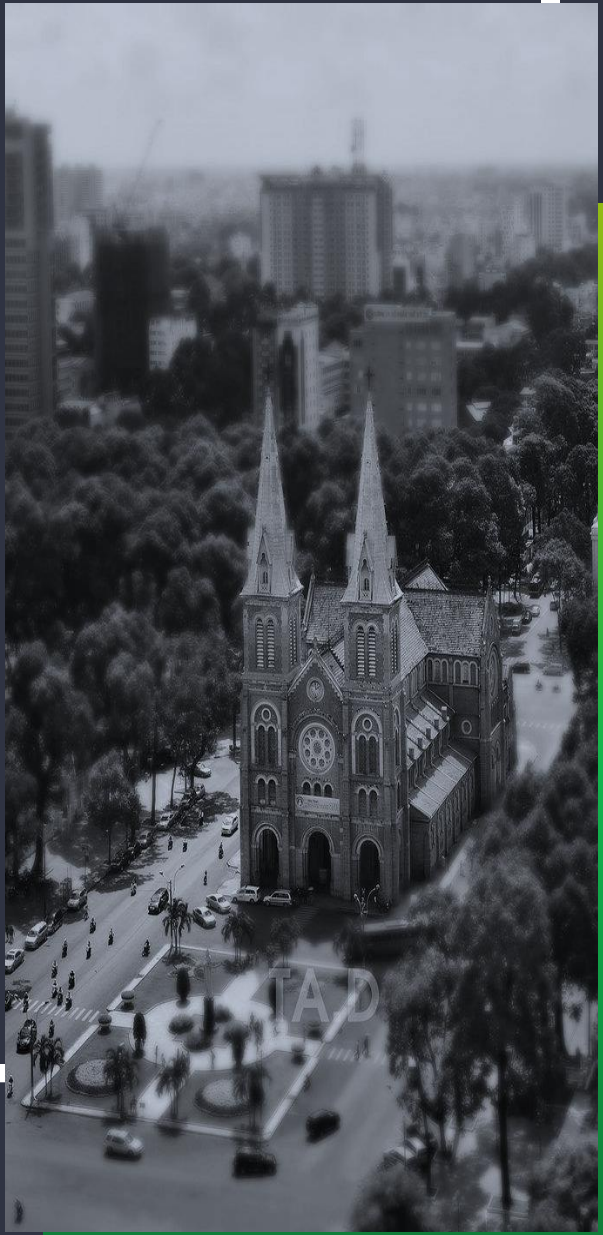
BUSINESS PROBLEM

THE PURPOSE OF THIS PROJECT

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

- Investors who are interested in opening a business in Hanoi, Vietnam, especially foreigners.
- Local authorities could 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
- Other stakeholders



DATA SECTION

- A list of all neighborhoods in Hanoi: from Wikipedia with BeautifulSoup: https://en.wikipedia.org/wiki/Category:Districts_of_Hanoi
- The coordinate data of Hanoi city, including specific latitude and longitude of all neighborhoods: extracted by using Geocoder package
- Data of venues around the neighborhoods in Hanoi: retrieved by using Foursquare API

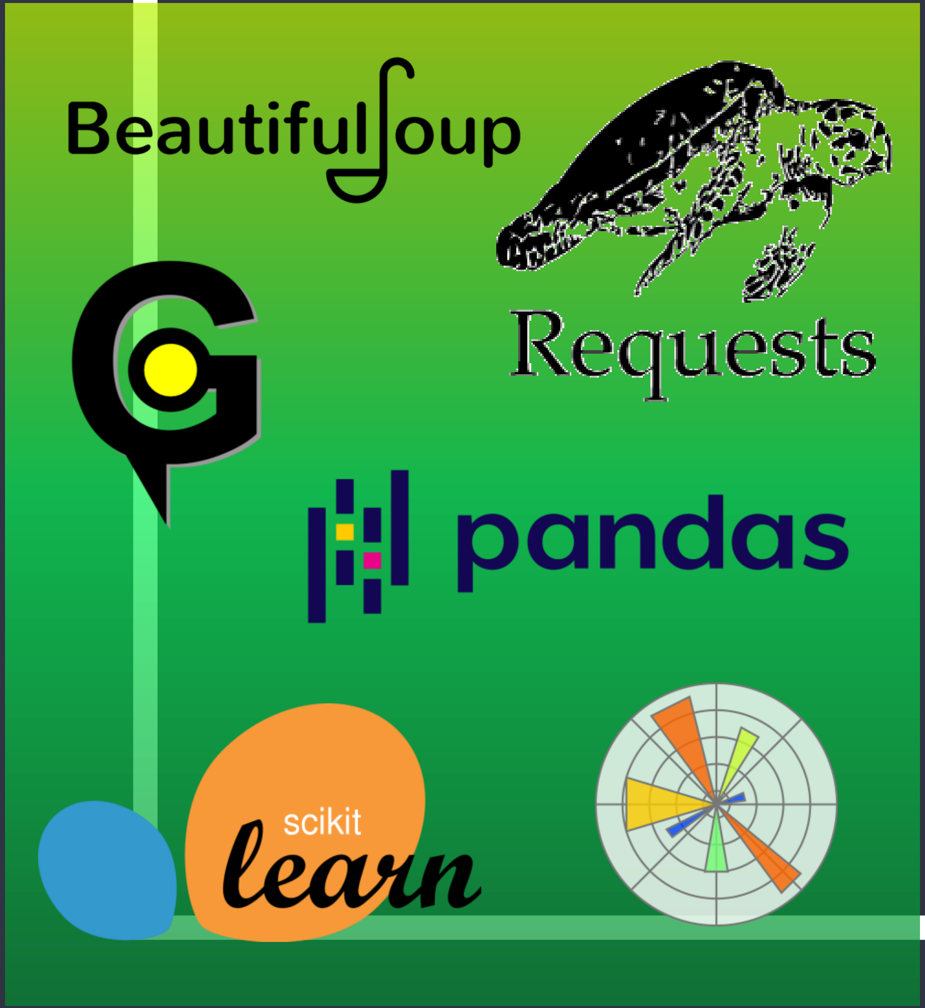
2

METHODOLOGY

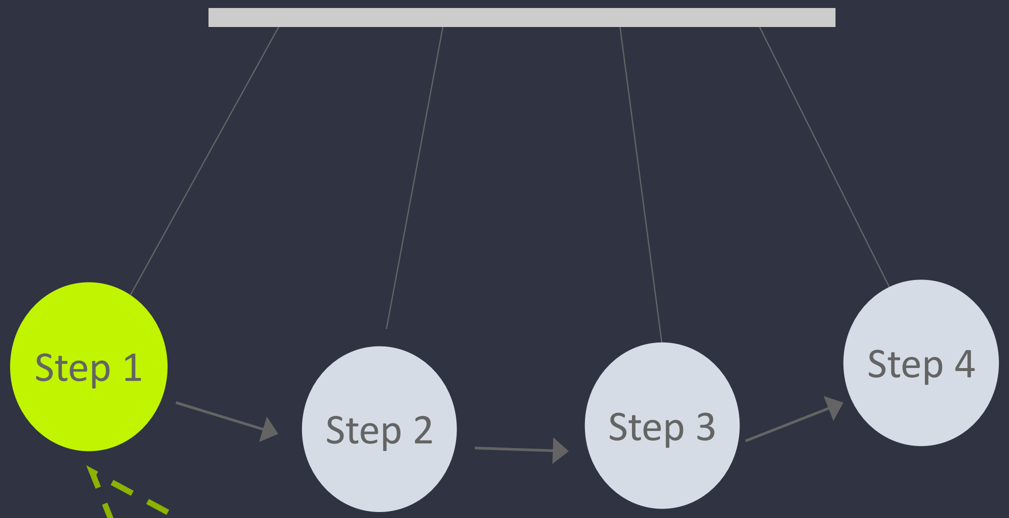
Methodology in the capstone project

PACKAGES AND LIBRARIES

- 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
- Foilum: map rendering library
- Kmeans from sklearn: used for clustering venues



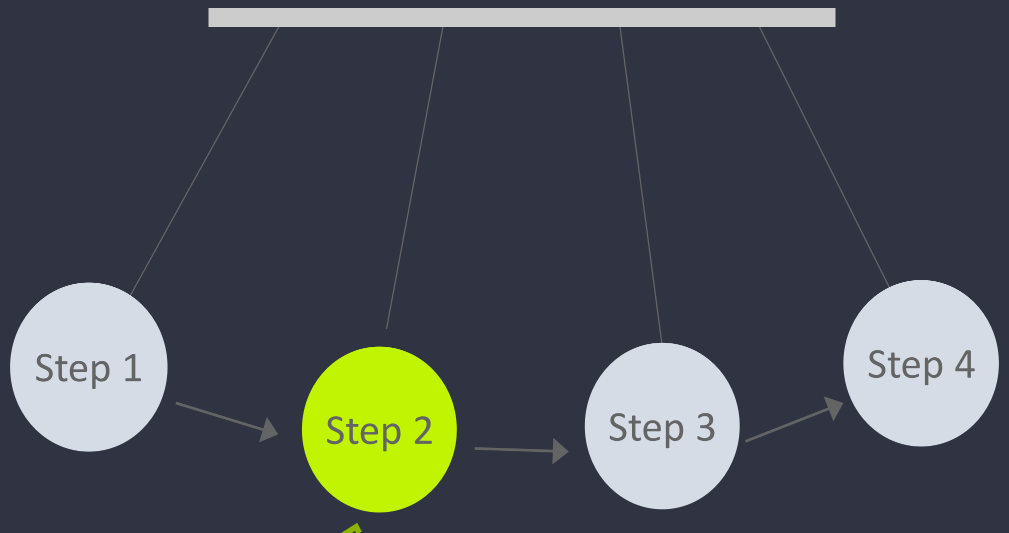
STEP BY STEP TO THE RESULT



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.

	Neighborhood
0	Ba Đình
1	Ba Vì
2	Bắc Từ Liêm
3	Cầu Giấy
4	Chương Mỹ
5	Đan Phượng
6	Đông Anh
7	Đống Đa
8	Gia Lâm
9	Hà Đông

STEP BY STEP TO THE RESULT



Step 1

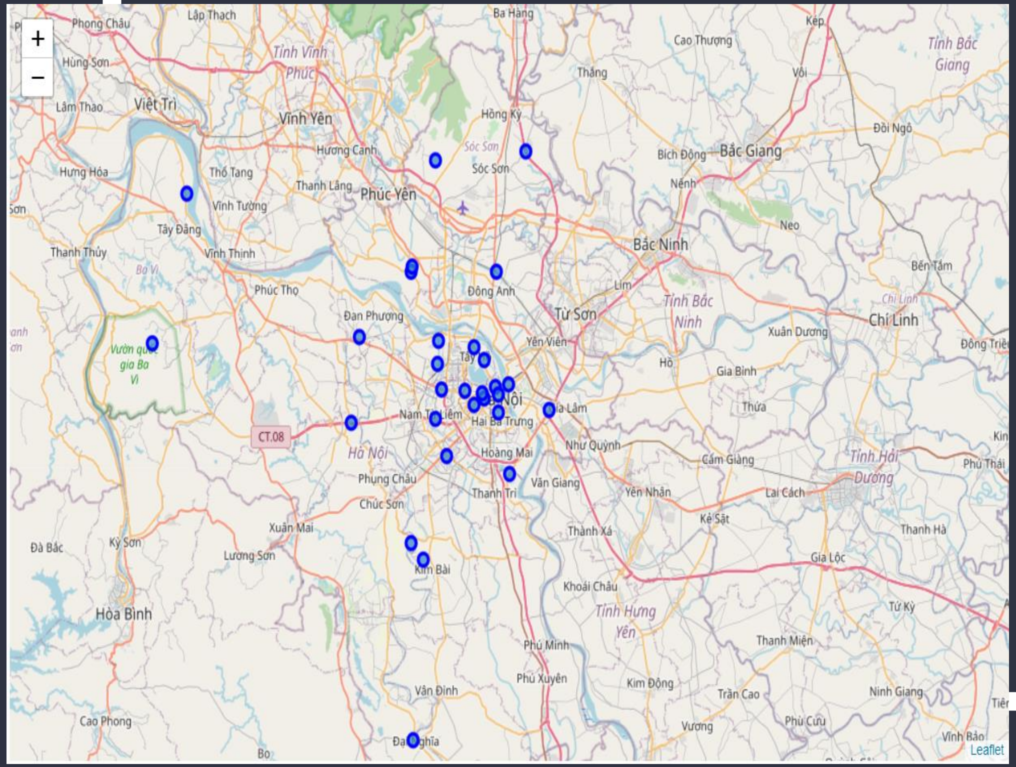
Step 2

Step 3

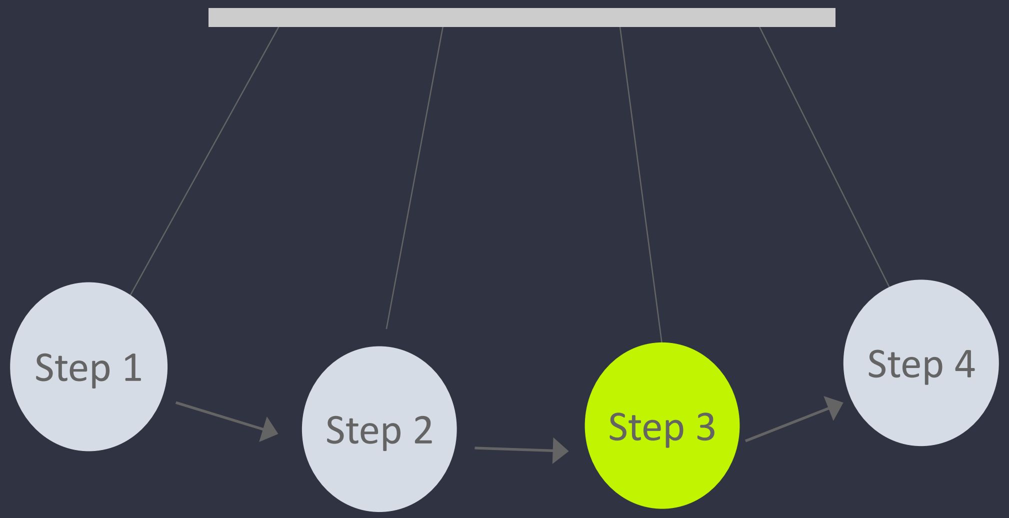
Step 4

- ○ Pull off latitude and longitude of each neighborhood with Geocoder package.
- ○ Add them as new columns to the previous Dataframe.
- ○ Plot a map of Hanoi city with each neighborhood is presented as a marker

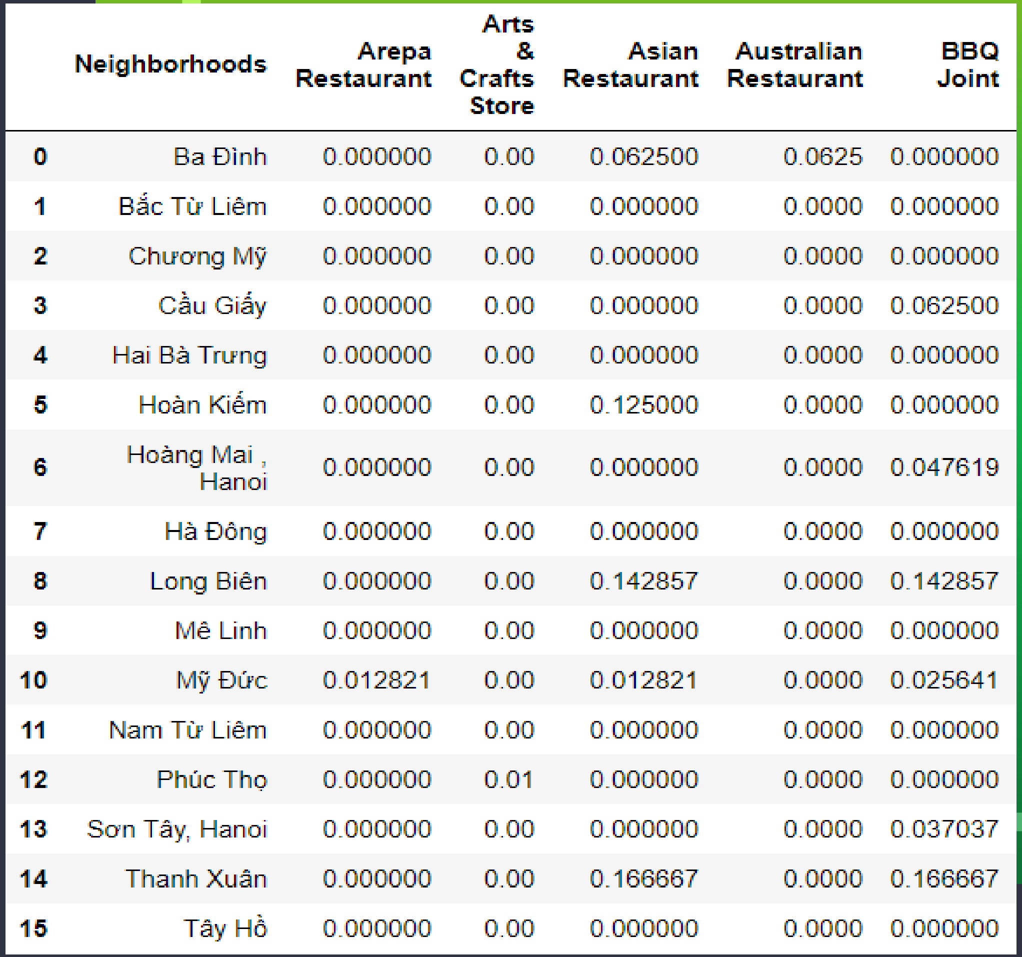
Most of neighborhoods are distributed around the center
=> There is a possibility that this area could be the ideal spot to attract more people



STEP BY STEP TO THE RESULT

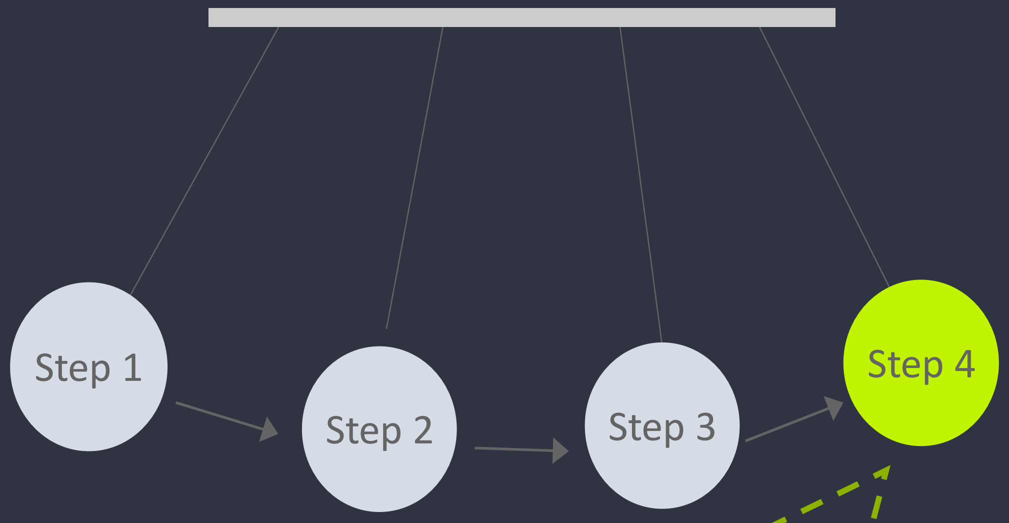


- Get the top 100 venues (Venue name, Venue Latitude, Venue Longitude and Venue Category) that are within a radius of 500 meters by Foursquare API.
- Check how many venues returned for each Neighborhood and find out how many unique categories can be curated from all the returned venues by using pandas
- Group rows by neighborhood and taking the mean of the frequency of occurrence of each category

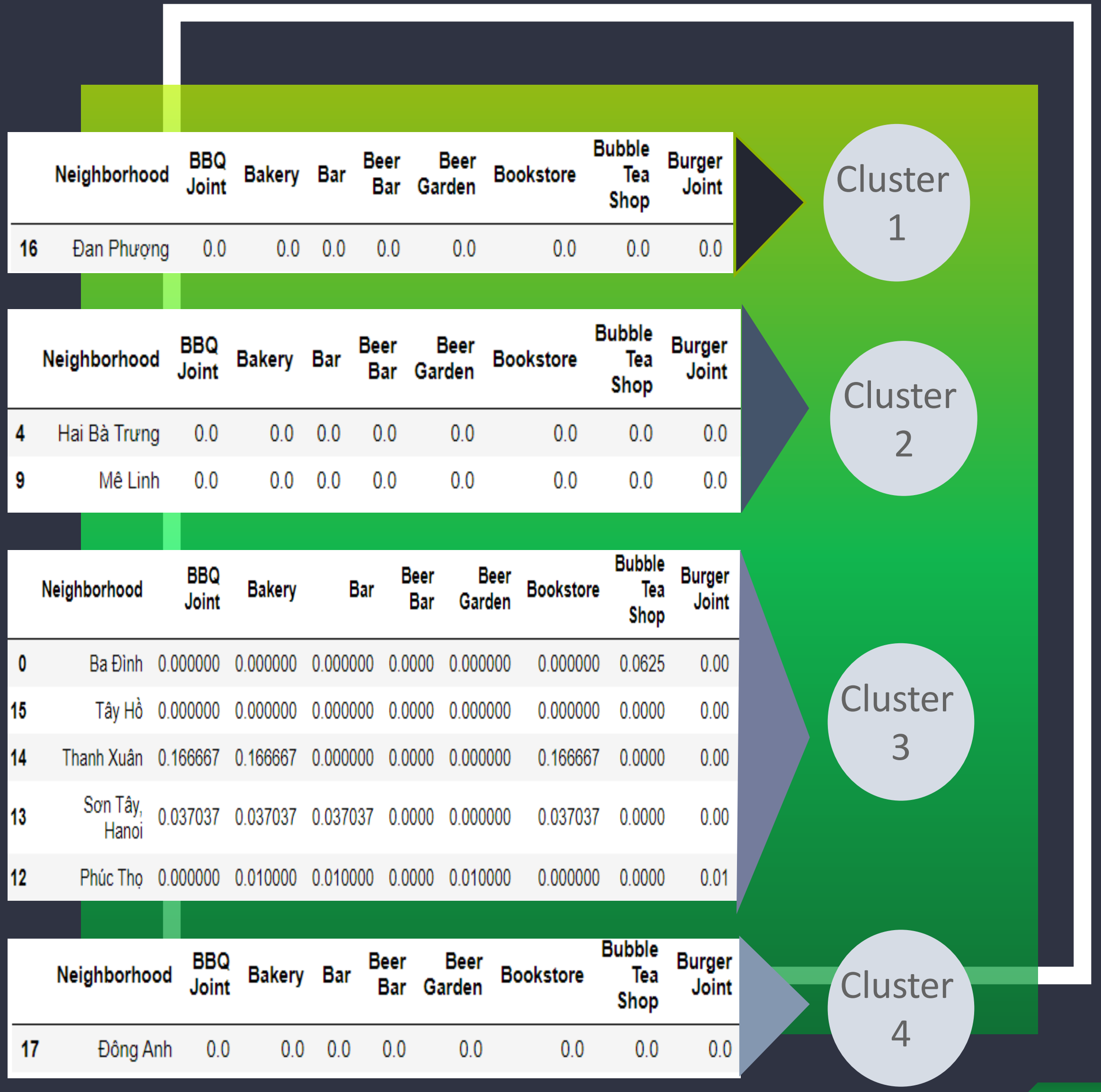


	Neighborhoods	Arepa Restaurant	Arts & Crafts Store	Asian Restaurant	Australian Restaurant	BBQ Joint
0	Ba Đình	0.000000	0.00	0.062500	0.0625	0.000000
1	Bắc Từ Liêm	0.000000	0.00	0.000000	0.0000	0.000000
2	Chương Mỹ	0.000000	0.00	0.000000	0.0000	0.000000
3	Cầu Giấy	0.000000	0.00	0.000000	0.0000	0.062500
4	Hai Bà Trưng	0.000000	0.00	0.000000	0.0000	0.000000
5	Hoàn Kiếm	0.000000	0.00	0.125000	0.0000	0.000000
6	Hoàng Mai , Hanoi	0.000000	0.00	0.000000	0.0000	0.047619
7	Hà Đông	0.000000	0.00	0.000000	0.0000	0.000000
8	Long Biên	0.000000	0.00	0.142857	0.0000	0.142857
9	Mê Linh	0.000000	0.00	0.000000	0.0000	0.000000
10	Mỹ Đức	0.012821	0.00	0.012821	0.0000	0.025641
11	Nam Từ Liêm	0.000000	0.00	0.000000	0.0000	0.000000
12	Phúc Thọ	0.000000	0.01	0.000000	0.0000	0.000000
13	Sơn Tây, Hanoi	0.000000	0.00	0.000000	0.0000	0.037037
14	Thanh Xuân	0.000000	0.00	0.166667	0.0000	0.166667
15	Tây Hồ	0.000000	0.00	0.000000	0.0000	0.000000

STEP BY STEP TO THE RESULT



Finally, based on the above cleaned data, we will run k-means to cluster Hanoi into 4 clusters. => The results will allow us to identify which neighborhoods have higher concentration of businesses, and help us determine which area is the best location to open a business in Hanoi



	Neighborhood	BBQ Joint	Bakery	Bar	Beer Bar	Beer Garden	Bookstore	Bubble Tea Shop	Burger Joint
16	Đan Phượng	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Cluster 1

	Neighborhood	BBQ Joint	Bakery	Bar	Beer Bar	Beer Garden	Bookstore	Bubble Tea Shop	Burger Joint
4	Hai Bà Trưng	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9	Mê Linh	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Cluster 2

	Neighborhood	BBQ Joint	Bakery	Bar	Beer Bar	Beer Garden	Bookstore	Bubble Tea Shop	Burger Joint
0	Ba Đình	0.000000	0.000000	0.000000	0.0000	0.000000	0.000000	0.0625	0.00
15	Tây Hồ	0.000000	0.000000	0.000000	0.0000	0.000000	0.000000	0.0000	0.00
14	Thanh Xuân	0.166667	0.166667	0.000000	0.0000	0.000000	0.166667	0.0000	0.00
13	Sơn Tây, Hanoi	0.037037	0.037037	0.037037	0.0000	0.000000	0.037037	0.0000	0.00
12	Phúc Thọ	0.000000	0.010000	0.010000	0.0000	0.010000	0.000000	0.0000	0.01

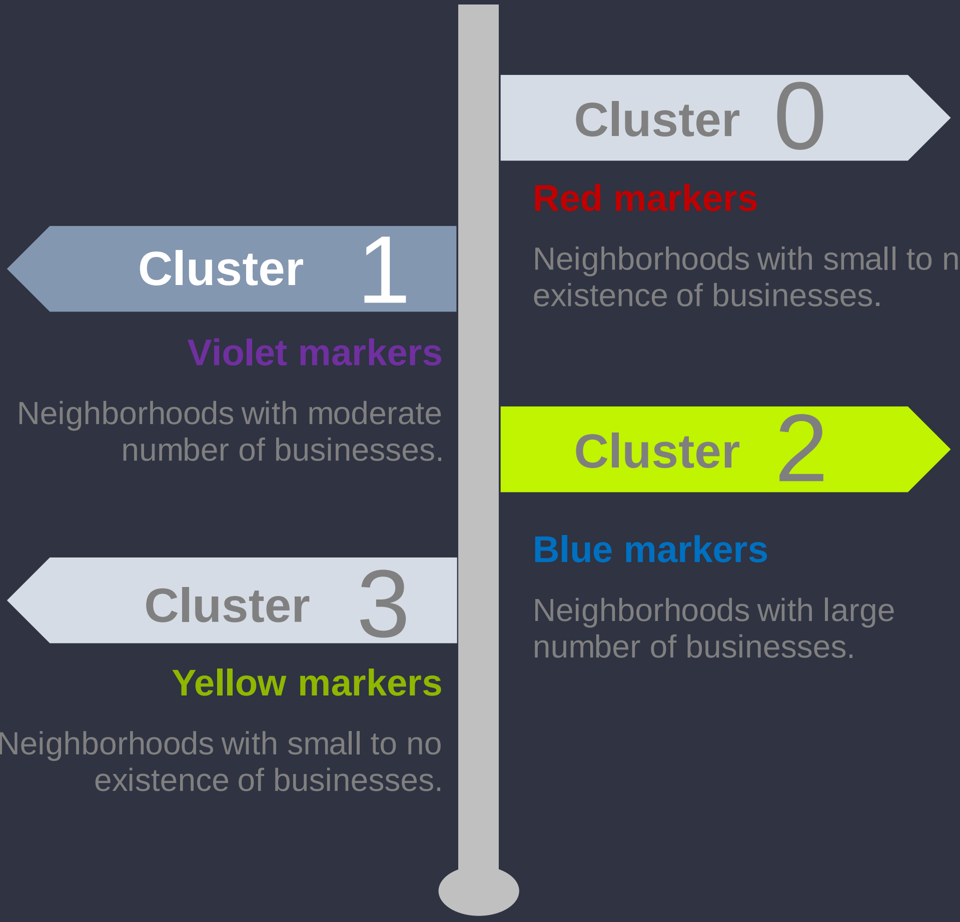
	Neighborhood	BBQ Joint	Bakery	Bar	Beer Bar	Beer Garden	Bookstore	Bubble Tea Shop	Burger Joint
17	Đông Anh	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

3

RESULTS

Results of capstone project

RESULTS



Cluster 0

Red markers

Neighborhoods with small to no existence of businesses.

Cluster 1

Violet markers

Neighborhoods with moderate number of businesses.

Cluster 2

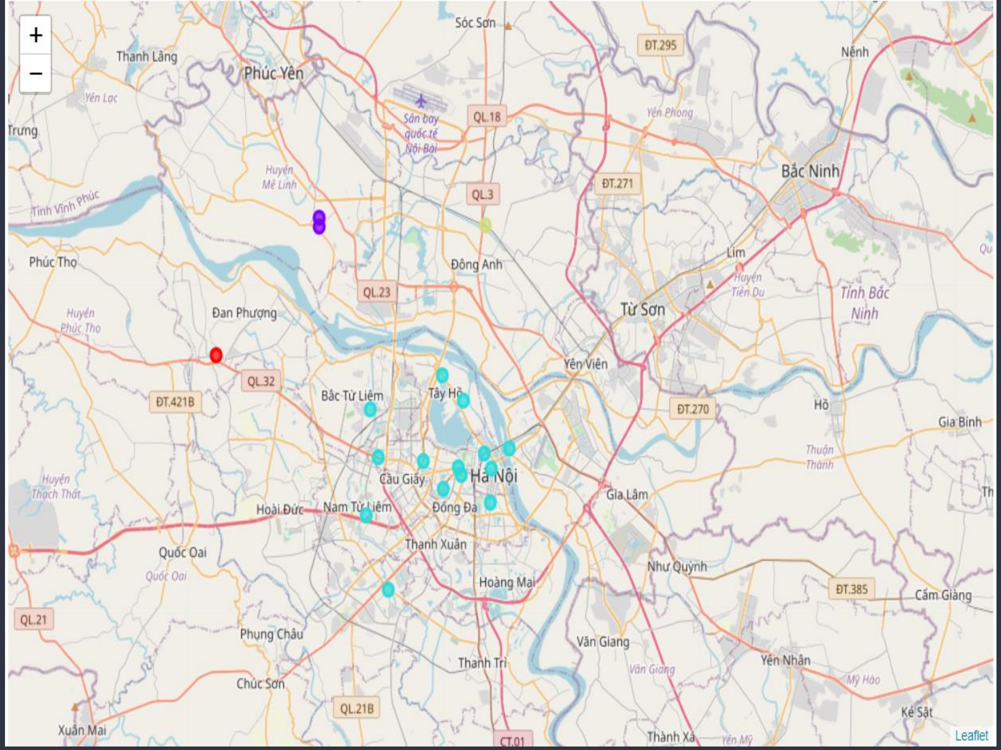
Blue markers

Neighborhoods with large number of businesses.

Cluster 3

Yellow markers

Neighborhoods with small to no existence of businesses.



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

4

SUMMARY

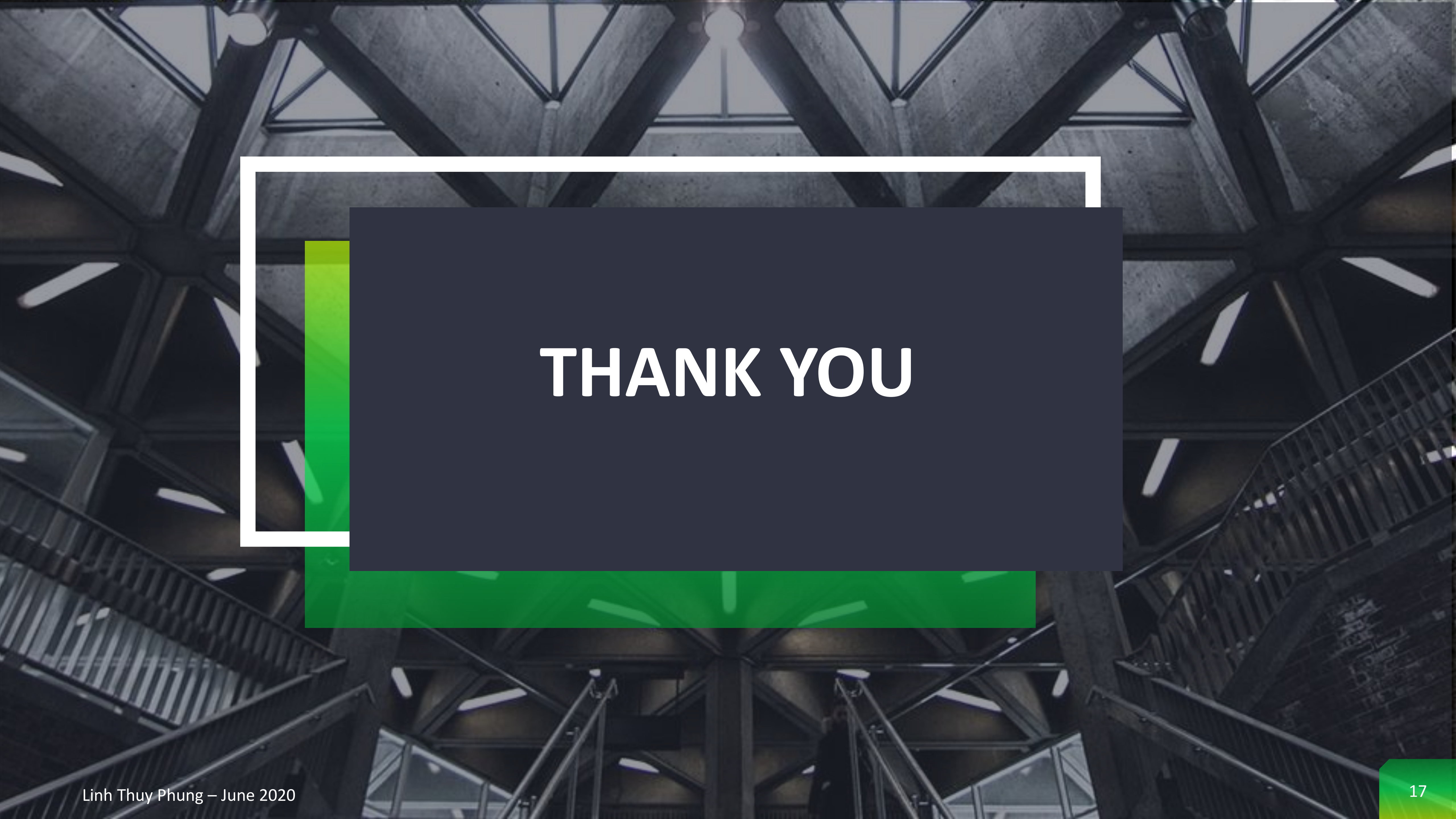
Discussion and Conclusion

DISCUSS

According to the results, the cluster 2 will be the most suitable location to open a new business, because it has the highest frequency of occurrence of businesses. Normally, people want to avoid areas which are already populated due to high competition. However, it is not the case of Hanoi, instead, investors should do more research to find the unique business niche, since Hanoi is a rich cultural capital city, with many traditional jobs. Meanwhile, 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

CONCLUDE

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. From there, we suggest that cluster 2 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 to place factories or plant natural resources. In the end, the project only gives an overview of the area to suggest the most suitable location, but neglecting many other factors. Therefore, future research based on this result should be conducted with more factors involved to overcome limitations of this project



THANK YOU