Applied Data Science Capstone

The Battle of Neighborhoods

IBM Data Science Professional Certificate

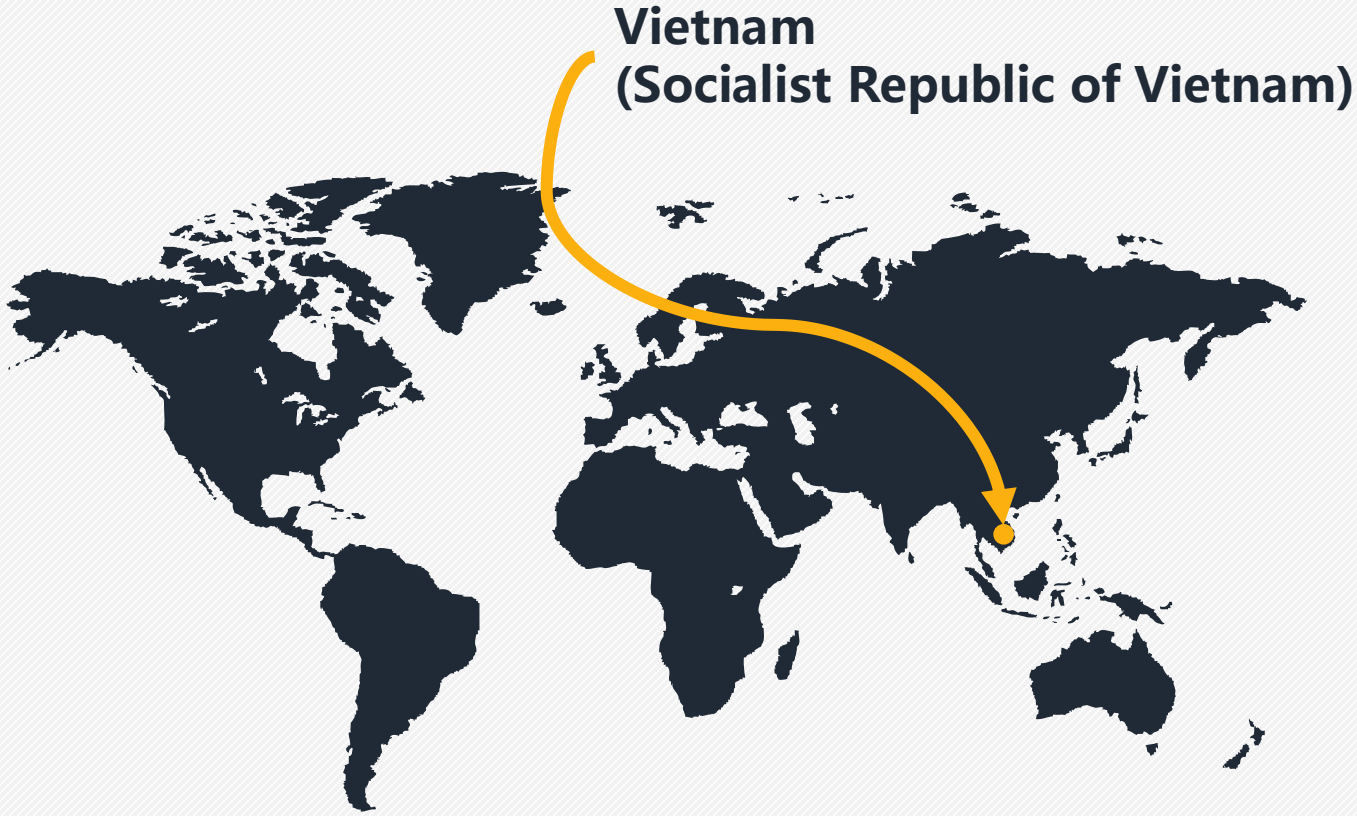
By: Mak Chun Wai, Michael
Finishing date: 15/05/2021

CONTENTS

Choice of country

Southeast Asia

At the eastern edge of the Indochinese Peninsula



58 provinces and 5 municipalities

Covering area 331,699 km^2

Population: >96 million inhabitants

16th most populous country

Background of the project

Source: The World Bank
https://data.worldbank.org/indicator/NY.GDP.PCAP.KD.ZG?end=2019&locations=VN&start=2009

Introduction

As one of the rising stars among the developing countries, Vietnam is a popular target for investors. This project is to try to cluster the provinces/cities in Vietnam and find out the best business chance by considering different clusters

Target audience

The result of this project/study is targeting to the individuals/groups who would like to own/run a business in Vietnam without any limitation in choosing the category of business

Data: web scraping and Foursquare

Wikipedia

List of cities/provinces in Vietnam: https://en.wikipedia.org/wiki/Postal_codes_in_Vietnam
Latitude and Longitude: https://geohack.toolforge.org/ (in each province's Wikipedia link)
GDP per capita (2011 PPP US$): https://en.wikipedia.org/wiki/Provinces_of_Vietnam

Foursquare API

Top venues within certain radius of each cities/provinces
(only used venue latitude, longitude and category in this project)

Data cleansing

Wikipedia

From the list of cities/provinces in Vietnam, 68 cities/provinces are obtained
Latitude and longitude of all cities/provinces are scraped and added into the same dataframe
After merging the dataframe with the list of GDP per capita, 61 cities/provinces are left

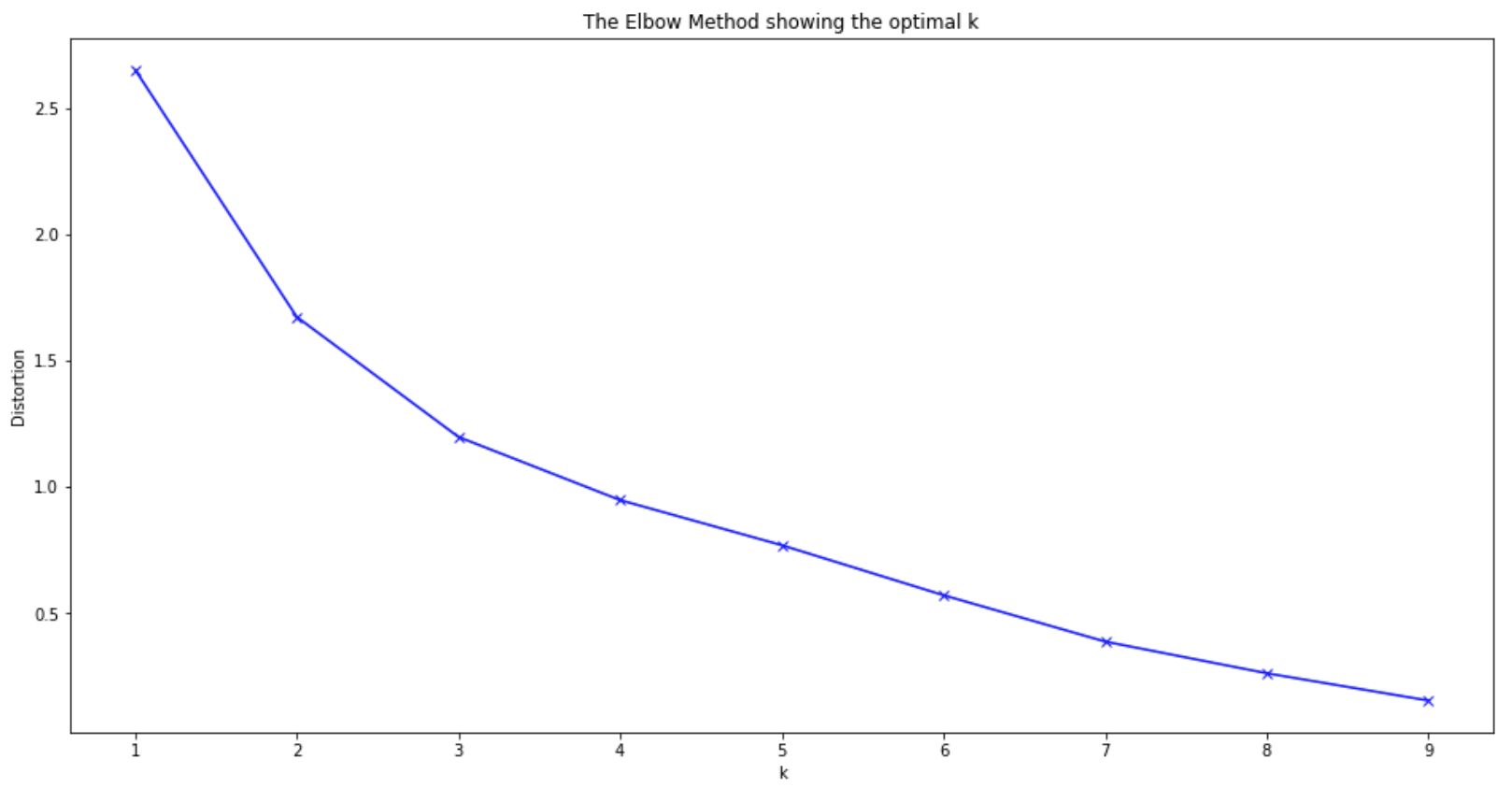
Foursquare API

By using the API, top 100 venues within a radius of 500 meters to all cities/provinces are obtained.
And surprisingly, only 167 venues with 56 unique categories are found in total.
They are in 12 cities/provinces.
(Possible reason: data in Foursquare may not cover so many areas in Vietnam)

Unsupervised learning: k-means clustering

Define k

We can use the elbow method introduced in the course of Machine Learning with Python to find the optimal k, 3 in this case



Result of clustering

Cluster 1

	Area	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	PostalCode	Latitude	Longitude	GDP per capita in USD
11	Đắk Lắk Province	Food	Vietnamese Restaurant	Dessert Shop	Gym / Fitness Center	French Restaurant	Food Truck	Food Court	Food & Drink Shop	Flea Market	Fast Food Restaurant	63000	12.666667	108.05	2,555.78

Result of clustering

Cluster 2

	Area	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	PostalCode	Latitude	Longitude	GDP per capita in USD
1	Cần Thơ	Coffee Shop	Hotel	Café	Vietnamese Restaurant	Bed & Breakfast	Multiplex	Park	Bakery	Asian Restaurant	Bar	94000	10.033333	105.783333	6,260.52
2	Da Nang	Vietnamese Restaurant	Train Station	Shopping Mall	Fast Food Restaurant	Burger Joint	Café	Pizza Place	Movie Theater	Dim Sum Restaurant	Food Truck	50000	16.069444	108.209722	4,811.58
3	Hai Phong	Vietnamese Restaurant	Restaurant	Café	Dessert Shop	Gym / Fitness Center	French Restaurant	Food Truck	Food Court	Food & Drink Shop	Food	04000	20.865139	106.683833	3,849.18
4	Hanoi	Coffee Shop	Café	Hotel	Vietnamese Restaurant	Noodle House	Hotel Bar	Ice Cream Shop	Italian Restaurant	Sushi Restaurant	Steakhouse	10000	21.028333	105.854167	3,923.21
5	Ho Chi Minh City	Noodle House	Seafood Restaurant	Flea Market	Fast Food Restaurant	Multiplex	Pizza Place	Gym / Fitness Center	BBQ Joint	Sushi Restaurant	Supermarket	70000	10.8	106.65	7,147.09
6	Hà Tĩnh Province	Vietnamese Restaurant	Asian Restaurant	Hotel	Dim Sum Restaurant	Gym / Fitness Center	French Restaurant	Food Truck	Food Court	Food & Drink Shop	Food	45000	18.333333	105.9	2,117.32
7	Hậu Giang Province	Vietnamese Restaurant	Supermarket	Shopping Plaza	Coffee Shop	Australian Restaurant	BBQ Joint	French Restaurant	Food Truck	Food Court	Food & Drink Shop	95000	9.783333	105.466667	2,943.84
8	Khánh Hòa Province	Restaurant	Hotel	Music Venue	Gym / Fitness Center	Coffee Shop	Café	Bowling Alley	Lounge	Bakery	Food Court	57000	12.25	109.2	4,180.68
9	Lâm Đồng Province	Vietnamese Restaurant	Hotel	Ice Cream Shop	Café	Dim Sum Restaurant	Hostel	Seafood Restaurant	Food Truck	Food Court	Food & Drink Shop	66000	11.95	108.433333	3,331.80
10	Điện Biên Province	Historic Site	Vietnamese Restaurant	Hotel	History Museum	Steakhouse	Cultural Center	Food Truck	Food Court	Food & Drink Shop	Food	32000	21.383333	103.016667	1,589.03

Result of clustering

Cluster 3

	Area	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	PostalCode	Latitude	Longitude	GDP per capita in USD
0	Bình Thuận Province	Asian Restaurant	Food Truck	Vietnamese Restaurant	Dessert Shop	Gym / Fitness Center	French Restaurant	Food Court	Food & Drink Shop	Food	Flea Market	77000	10.933333	108.1	3,090.17

Discussion

GDP per capita

I have scraped the GDP per capita to represent the wealth level of each area. As a businessman, to earn more profit, it would be a better choice to develop the business in an area which the citizens are richer so that their buying power and probably their behavior and habits on spending will contribute the company's profit

Choice/chance for business

Based on the result of the clustering, I will make some assumptions so that conclusion could be drawn from them. The details would be mentioned in the conclusion section

Conclusion: deciding the business

	Area	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	PostalCode	Latitude	Longitude	GDP per capita in USD
1	Cần Thơ	Coffee Shop	Hotel	Café	Vietnamese Restaurant	Bed & Breakfast	Multiplex	Park	Bakery	Asian Restaurant	Bar	94000	10.033333	105.783333	6,260.52
2	Da Nang	Vietnamese Restaurant	Train Station	Shopping Mall	Fast Food Restaurant	Burger Joint	Café	Pizza Place	Movie Theater	Dim Sum Restaurant	Food Truck	50000	16.069444	108.209722	4,811.58
3	Hai Phong	Vietnamese Restaurant	Restaurant	Café	Dessert Shop	Gym / Fitness Center	French Restaurant	Food Truck	Food Court	Food & Drink Shop	Food	04000	20.865139	106.683833	3,849.18
4	Hanoi	Coffee Shop	Café	Hotel	Vietnamese Restaurant	Noodle House	Hotel Bar	Ice Cream Shop	Italian Restaurant	Sushi Restaurant	Steakhouse	10000	21.028333	105.854167	3,923.21
5	Ho Chi Minh City	Noodle House	Seafood Restaurant	Flea Market	Fast Food Restaurant	Multiplex	Pizza Place	Gym / Fitness Center	BBQ Joint	Sushi Restaurant	Supermarket	70000	10.8	106.65	7,147.09
6	Hà Tĩnh Province	Vietnamese Restaurant	Asian Restaurant	Hotel	Dim Sum Restaurant	Gym / Fitness Center	French Restaurant	Food Truck	Food Court	Food & Drink Shop	Food	45000	18.333333	105.9	2,117.32
7	Hậu Giang Province	Vietnamese Restaurant	Supermarket	Shopping Plaza	Coffee Shop	Australian Restaurant	BBQ Joint	French Restaurant	Food Truck	Food Court	Food & Drink Shop	95000	9.783333	105.466667	2,943.84
8	Khánh Hòa Province	Restaurant	Hotel	Music Venue	Gym / Fitness Center	Coffee Shop	Café	Bowling Alley	Lounge	Bakery	Food Court	57000	12.25	109.2	4,180.68
9	Lâm Đồng Province	Vietnamese Restaurant	Hotel	Ice Cream Shop	Café	Dim Sum Restaurant	Hostel	Seafood Restaurant	Food Truck	Food Court	Food & Drink Shop	66000	11.95	108.433333	3,331.80
10	Điện Biên Province	Historic Site	Vietnamese Restaurant	Hotel	History Museum	Steakhouse	Cultural Center	Food Truck	Food Court	Food & Drink Shop	Food	32000	21.383333	103.016667	1,589.03

Conclusion: deciding the business

Assumption

As opening a business in a new country is risky, we assume that investors would make use of economies of scale to choose the most common business in a relatively developed area in Vietnam

Cluster

As cluster 1 & 3 have only 1 area respectively, cluster 2 should be the only choice

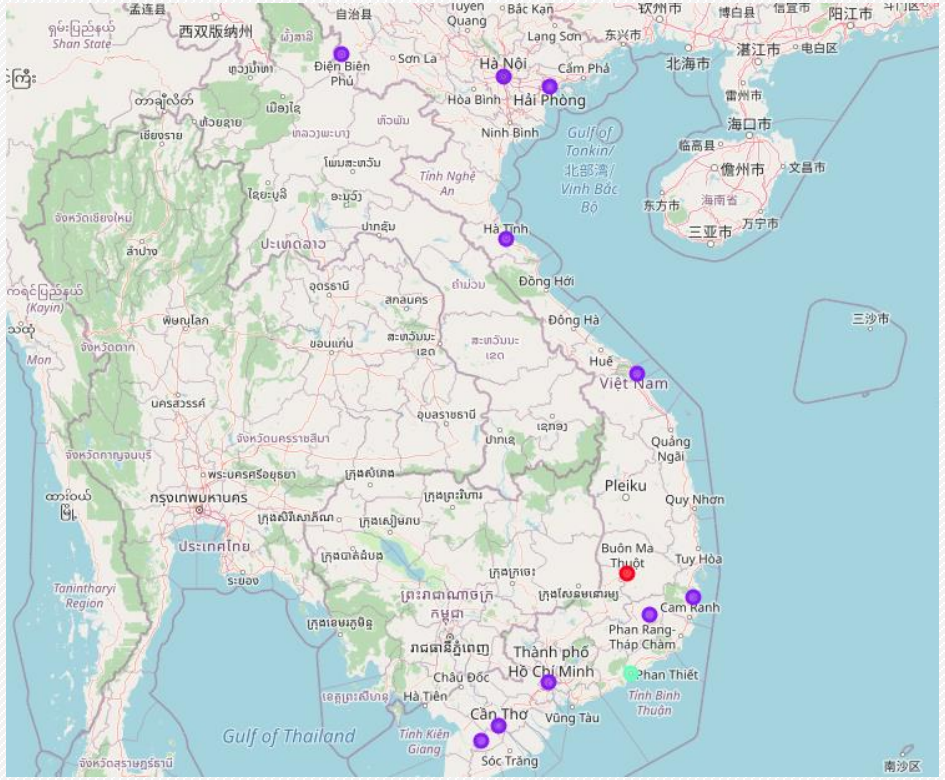
Business

In Cluster 2, Vietnamese Restaurant is the most common venue so it would be the safest choice under this assumption. If you need to make use of the effect of economies of scale, opening a Noodle House in Ho Chi Minh City or a Coffee Shop in Can Tho would be a good choice due to the high GDP in the area

Conclusion: factors of accuracy

Model for machine learning

I used k-means clustering in this project as it is the most common unsupervised learning model in machine learning. However, it may not be suitable for this topic. As you can observe from the map of Folium, the shape of Vietnam is relatively long and irregular, which is one of the drawbacks of k-means that does not work well with non-circular cluster shape



Conclusion: factors of accuracy

Availability of data

The data used may have large degree of bias due to unavailability. In the data provided by Foursquare, there are only 167 venues could be found in 12/60 areas in Vietnam, which is most likely not able to cover the major venues in Vietnam, and hence, create bias on the data and the result

```
In [20]: vietnam_grouped = vietnam_onehot.groupby('Area').mean().reset_index()
         vietnam_grouped
```

Out[20]:

	Area	Arts & Crafts Store	Asian Restaurant	Australian Restaurant	BBQ Joint	Bakery	Bar	Bed & Breakfast	Bowling Alley	Burger Joint	...	Seafood Restaurant	Shopping Mall	Shopping Plaza	Smoothie Shop	Spa	Steakhouse	Supermarket	Sushi Restaurant	Train Station	Vietnam Restaur
0	Binh Thuận Province	0.000000	0.500000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000(
1	Cần Thơ	0.000000	0.076923	0.000000	0.0	0.076923	0.000000	0.076923	0.000000	0.000000	...	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.076
2	Da Nang	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.111111	...	0.000000	0.111111	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.111111	0.222
3	Hai Phong	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.333
4	Hanoi	0.010989	0.000000	0.010989	0.0	0.010989	0.010989	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.00	0.010989	0.021978	0.021978	0.000000	0.032967	0.000000	0.098
5	Ho Chi Minh City	0.000000	0.000000	0.000000	0.1	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.100000	0.000000	0.00	0.000000	0.000000	0.000000	0.100000	0.100000	0.000000	0.000(

```
In [21]: vietnam_grouped.shape
```

```
Out[21]: (12, 56)
```

Applied Data Science Capstone

~ End ~

IBM Data Science Professional Certificate

By: Mak Chun Wai, Michael

Finishing date: 15/05/2021