**Final Coursera Capstone Project**

**Applied Data Science**

Minh Dao

University of People

# Introduction

Ho Chi Minh is the biggest city of Vietnam, which leads not only the economy but also the services. Therefore, there would be a very crowded place where the stores and services company stand next to each other. The total area of Ho Chi Minh city is around 2,061.2 km2 with the population is over 10mil people (9m as statistics, but there are more people without any identity). This city has 5 rural districts and 16 urban districts. So, it would be a little too small in case that I just investigate a small administrative area as districts.

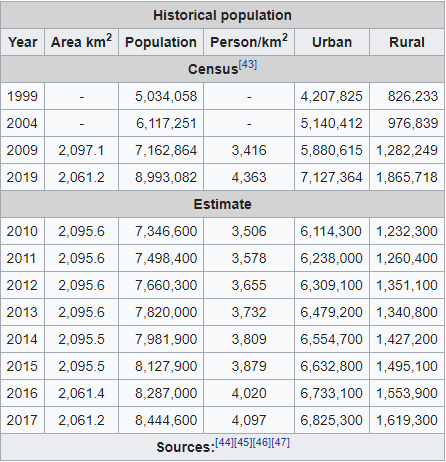


Figure 1: Demographic Statistics of Ho Chi Minh City

In terms of the leading economic engine with crowded population, it would be a fertile soil for any business want to challenge their creativity if they attend to high shares sectors, or if they want to trial new sectors. Hence, in this particular report, I want to challenge myself to investigate what model could I follow in term of a new business.

# Business Problem

There is a problem that as the biggest city in Vietnam, Ho Chi Minh city seems to have a large range of services as well as stores, or conveniences. It would be hard if you want to interfere into any existing fields. Furthermore, Ho Chi Minh city has developed for not a long time; its business on the large scale seems to focus into personnel store in stead of large business with multifactor combination.

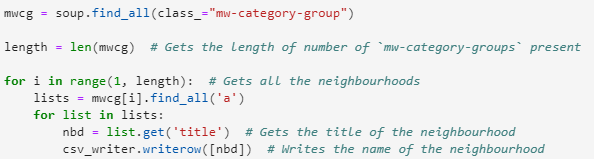
Other considerable factors are the traffic jam and population density. As you could look at the statistic above, the density is around 4,363 person/km2, which could easily lead to hard – to – move situations. This is the case when I have noticed that the operational range of each business would be small into the residential area, hardly expanding to the whole city.

# Data

The data for this project was gathered and analyzed from a variety of sources, with great care taken to ensure that the methods used were accurate.

# District

Web scraping with the BeautifulSoup library for Python can be used to retrieve data from Ho Chi Minh City's neighborhoods. The information about the neighborhood was taken from a Wikipedia article. You could access to the wiki page in ('https://en.wikipedia.org/wiki/Category:Districts\_of\_Ho\_Chi\_Minh\_City').

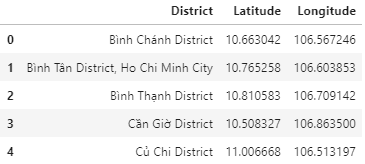


## Geocoding

The contents of the file hcmcity.csv are loaded into a Pandas Data Frame. The Google Maps Geocoding API is used to get the latitude and longitude of the Districts. The original dataframe is then modified with the geometric position values.

## Venue Data

The venue data is derived from the location data obtained after Web Scraping and Geocoding by transferring the necessary parameters to the FourSquare API and generating a new DataFrame that contains all of the venue information as well as the respective Districts.



# Methodology

A thorough analysis of the principles of methods, rules, and postulates employed g=have been made in order to ensure the inferences to be made are as accurate as possible.

## Accuracy of the Geocoding API

The number of incorrect results produced by the OpenCage Geocoder API was significant during the initial development process, prompting the creation of an algorithm to assess the accuracy of the Geocoding API in use. Geocoding APIs from different providers were checked as part of the algorithm, and Google Maps Geocoder API turned out to have the fewest collisions (errors) in our study.

## Folium

Folium is based on the Python ecosystem's data wrangling capabilities and the leaflet.js library's mapping capabilities. Folium is used to visualize all clusters, and then produces a Leaflet map based on OpenStreetMap technology.

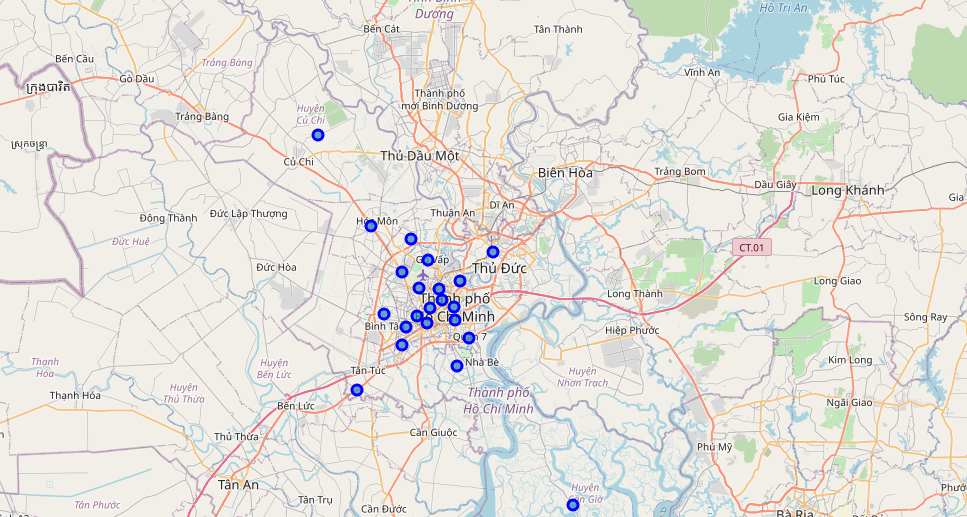


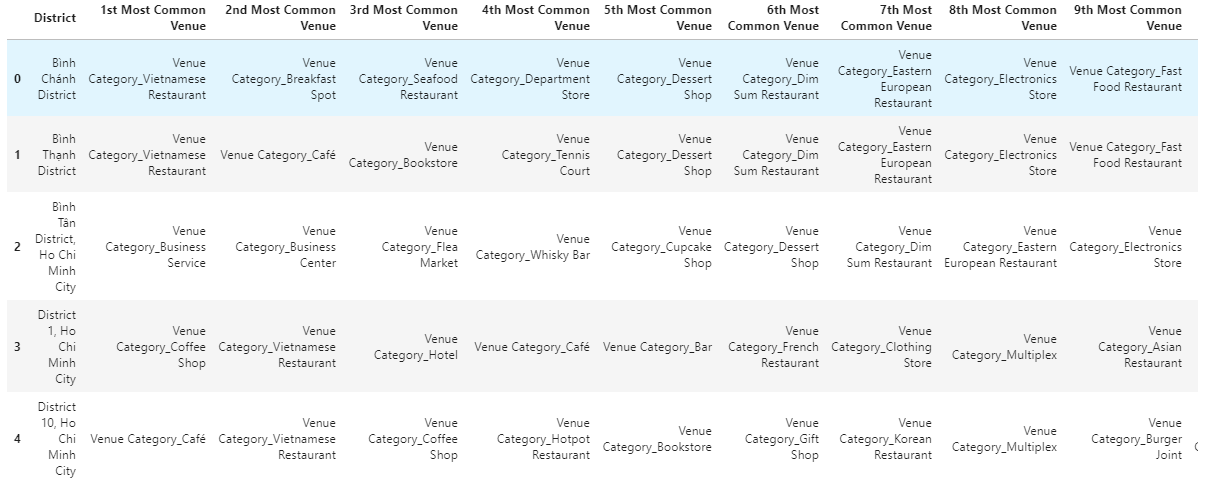
Figure 2: Districts of Ho Chi Minh City

## One hot encoding

One type of hot encoding is the conversion of categorical variables into a format that can be fed into machine learning algorithms to improve prediction accuracy. All specific things in the Venue Category are one-hot encoded for the K-means Clustering Algorithm.

## Top 10 most common venues

Because of the wide range of venues, only the top 10 are chosen, and a new DataFrame is created to train the K-means Clustering Algorithm.



## Optimal number of clusters

As compared to other clusters, the Silhouette Score is a measure of how close an object is to its own cluster (cohesion) (separation). The silhouette has a number of values ranging from -1 to +1, with a high value indicating that the object is well matched to its own cluster but poorly matched to neighboring clusters. The optimum cluster size is calculated based on the Silhouette Score of various clusters below 20.

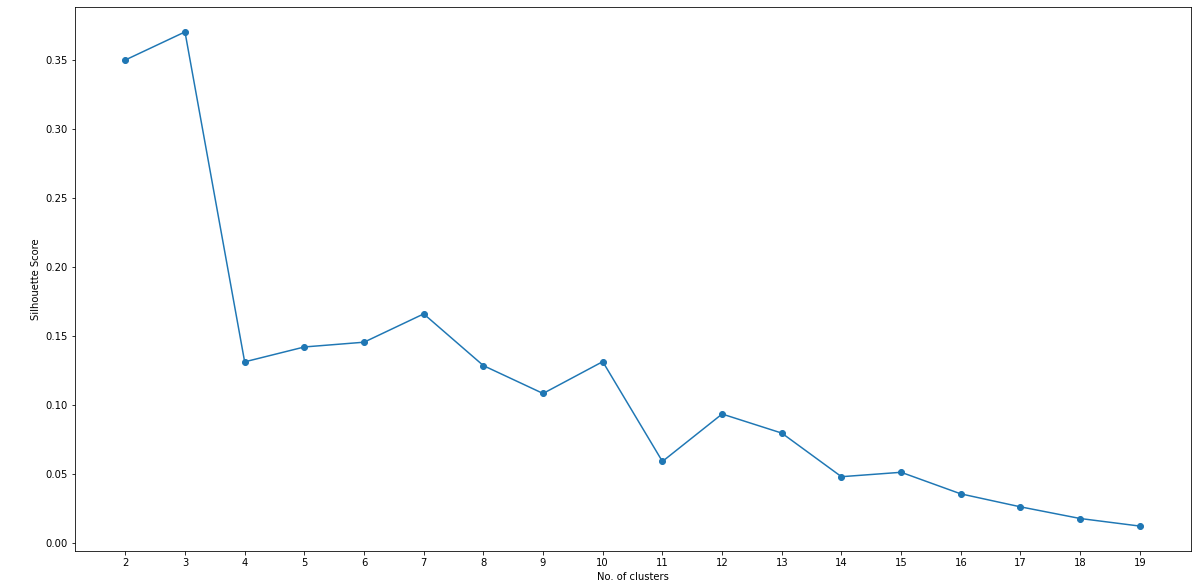
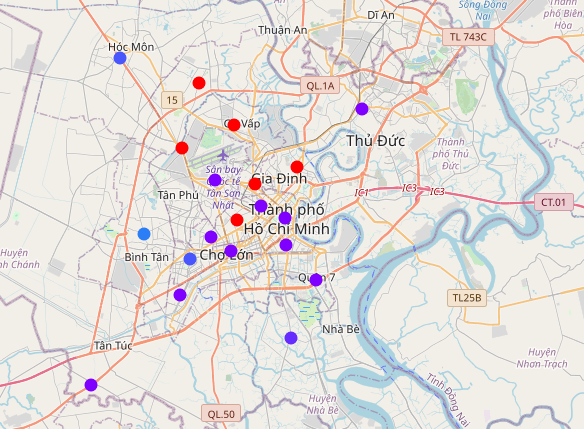
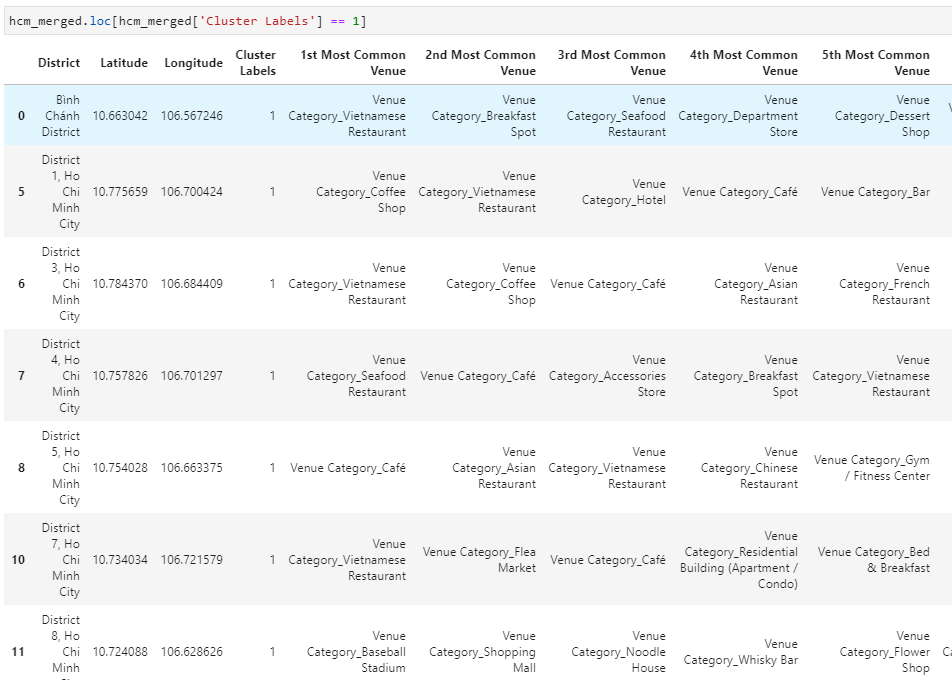


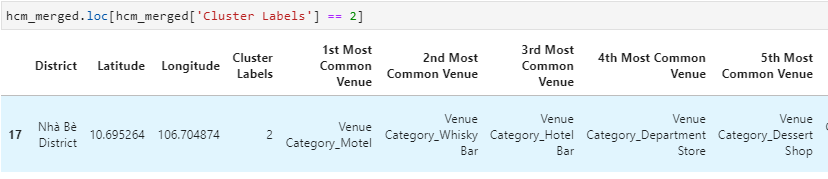
Figure 3: Silhouette score vs No.of clusters.

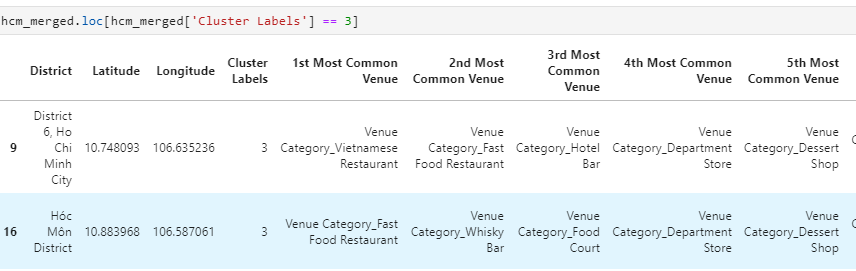
# Results

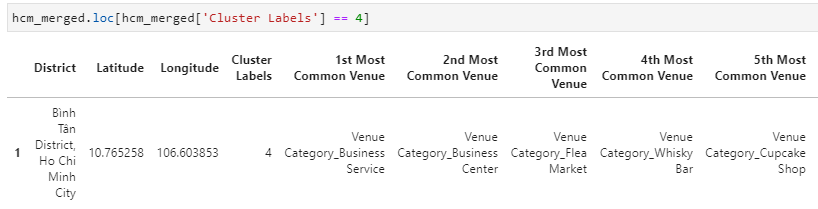
The districts are divided into n clusters, with n denoting the number of clusters discovered using the best method. Different colors are used to visualize the clustered districts so that they can be distinguished.





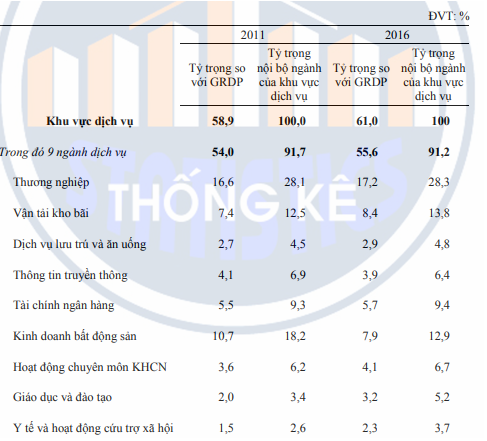






# Discussion

After analyzing the various clusters produced by the Machine learning algorithm, I feel that almost any district is sticked to the food and beverage industry. Hence, it would be extreme in case of any intention to step into this field. There are still chances for fitness service in large scale business with modern devices. As I investigate the whole sectors, they are almost focused in social services rather than personal service.



In circumstances of the hottest segments – food and beverage, I see the lack of variety in all clusters. It seems to be that many food stores are serving the same menus with less different, except for the center districts. Therefore, it would be chances if you could open some of the foreign menu restaurant in the suburb districts. I recommend to stay away to beverage store as the density of this is too high. The competitions would be too much intense.

As well as for the clothing stores, I notice that the density is high in the center districts, which means that many clothing stores are located here. It would be a good idea to move to surrounding district for more chance. However, people here tend to buy things in grouped stored. It means that despite of the high-density area, you have higher chance to approach potential customer

Moreover, there is shortage of complex area and the store for toys as well as gifts. This could be worth-investing sectors in term of acknowledging what could satisfy customer needs. The average income of this city is not so high, so they would be a little picky and confused when deciding to buy those things.

It would be a little strange when the electronics stores are not common. Maybe there is transition in buying habit from stores to online stores.

# Conclusion

As the biggest city in Vietnam, Ho Chi Minh City has attracted many people to come and settle down here. This is not only opportunity to come up with better life, but a fierce competition in earning money for living. More people, it means that more abilities that you could approach your potential customers. It is easy to see that almost the frequency in the data analysis turn to be food and beverage category. With the income in the middle level, people tend to focus most in the essentials for living, and eating come first in the list.

However, it could be a chance for you to expand in other categories, or in case that you could handle the issue of traffic jam and join the transportation industry.