Ho Chi Minh F&B Venues Location Analysis

# Business Problem

Ho Chi Minh city – Vietnam where I’m from is famous for the diverse cuisine and eating culture. You can find all kind of food & beverage (F&B) venues in this city, from street food stalls to fancy restaurants. As F&B plays an important role in Vietnamese’s daily life, running a restaurant is a very profitable business. However, due to the diversity of the city food culture, it is crucial to understand the overall picture of F&B venue types and how they are distributed & segmented before actually opening one.

In this project, I’ll address the following questions:

* What are the most frequently-visited F&B venues of Ho Chi Minh city?
* How are they distributed among the main districts and wards?
* How are they segmented base on the similarity of the most common F&B venues?
* What are the location and type of cuisine recommendation for opening a new restaurant?

# Who would be interested?

Start-ups who want to open a new restaurant.

Investors who want to look into F&B business.

City planners/ F&B Managers who need an overview on F&B venues distribution.

# Data

In this project I will use the following data source:

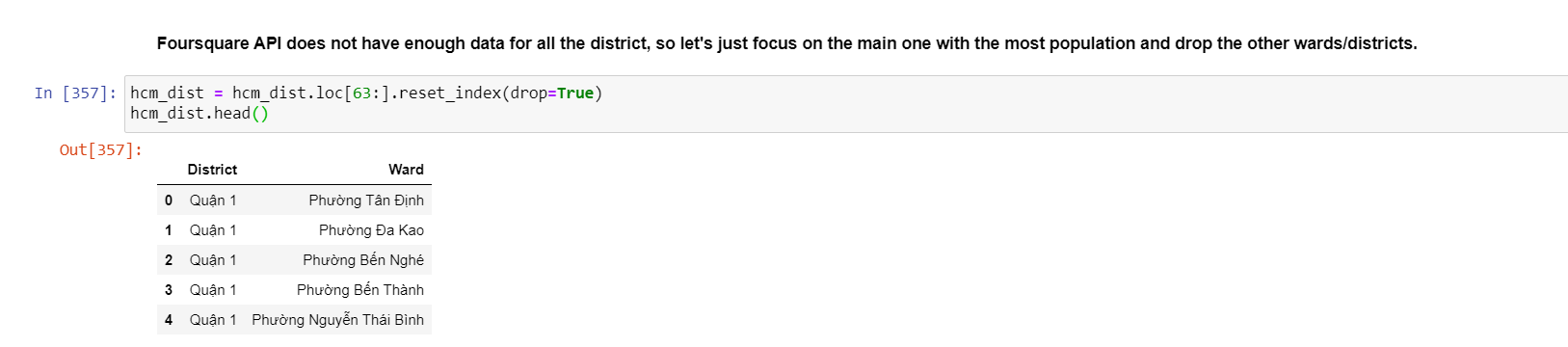
* List of districts/ wards of Ho Chi Minh city: scraping webpage: 'http://www.pso.hochiminhcity.gov.vn/web/guest/danhmucthongke-danhmucphuongxa'
* Location (latitude/longitude) of each ward: using Geocoder on each ward location.
* FourSquare API to generate nearby venues for each location. Which represent venues in that ward area. I used the special category ID for F&B venue.
* K-means Clustering were used to cluster all F&B venue base on its category. Each cluster is then analyzed for location recommendation.

# Methodology

## Ho Chi Minh District/Ward data - Web scraping

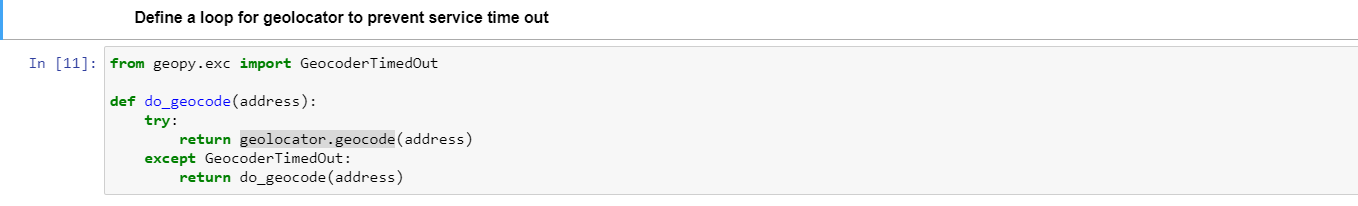
First I obtain the table of HCM districts and wards in this webpage: 'http://www.pso.hochiminhcity.gov.vn/web/guest/danhmucthongke-danhmucphuongxa'.

I’ve used read\_html instead of BeautifulSoup because it’s faster and more efficient. The list contain more than 24 districts, however not all districts data are available in Foursquare. Therefore to reduce computational error, I’ll just focus on 19 main districts and drop the rest.

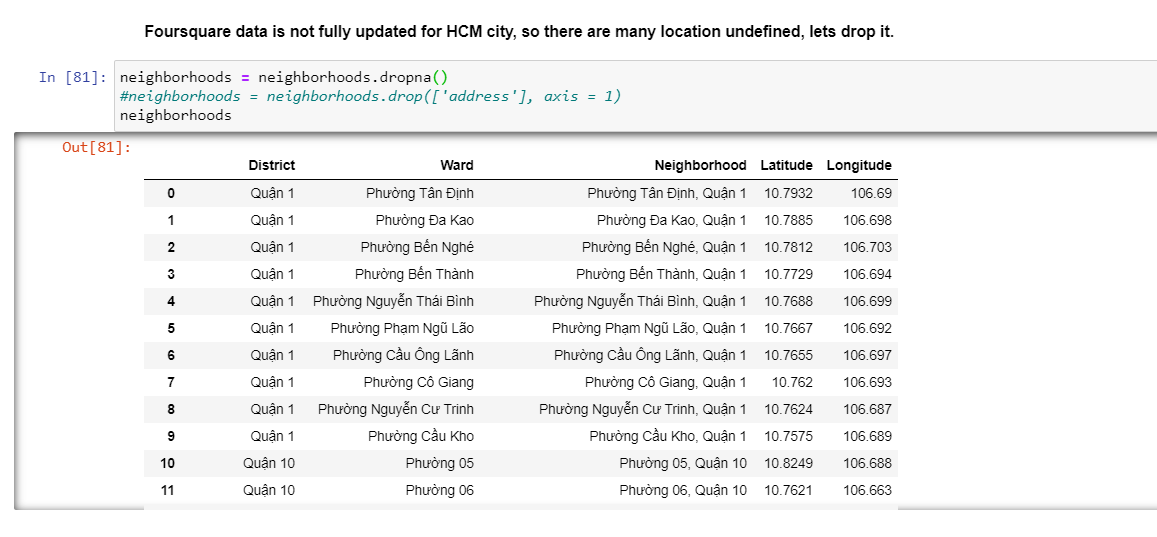


## Insert location data from geopy

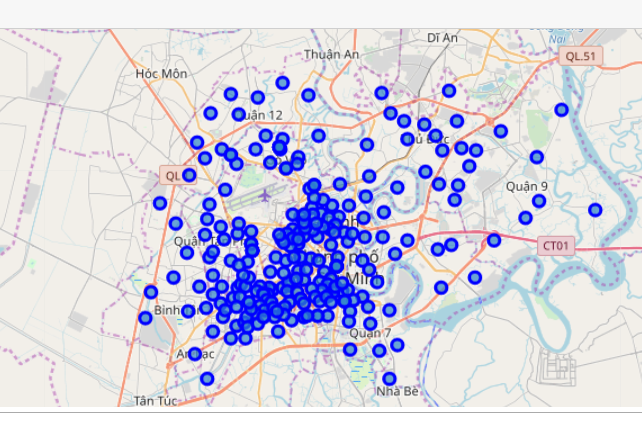
Next, I’ll get the coordinates of all total 196 wards/districts by apply the geolocator loop. However, the “geolocator.geocode” call sometimes result in “service timeout” as no data was returned from the server, resulted in breaking the loop. Therefore, I made another function to prevent this problem as follow:



Because FourSquare search engine is not sufficient for special language such as Vietnamese, several coordinates were not returned. I had to make some adjustment to the address name, both manually and by code loop. The final coordinates table is as follow:



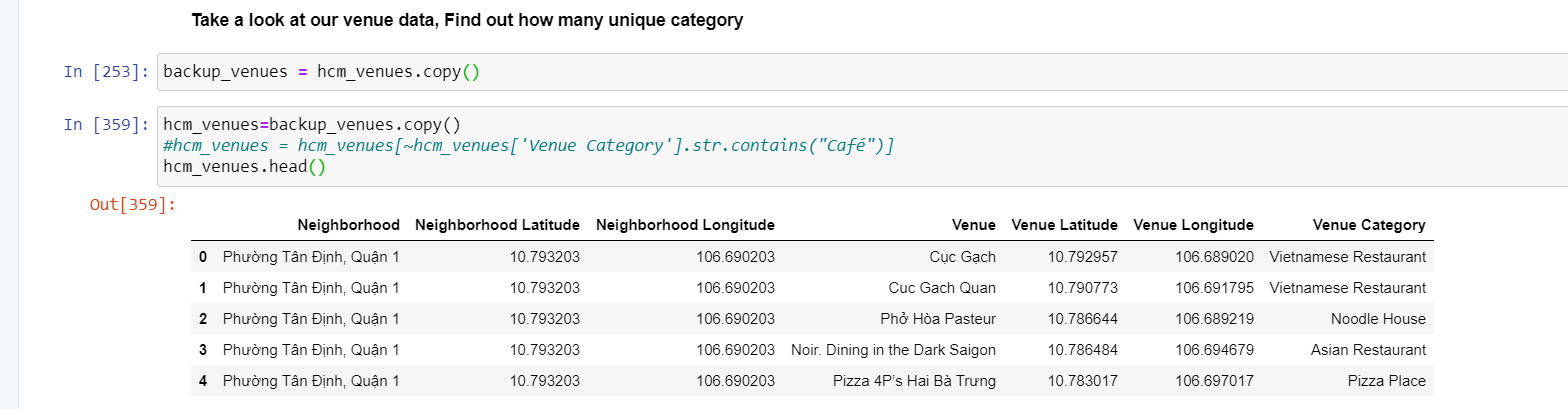
Let’s visualize the location of wards/districts of HCM city:



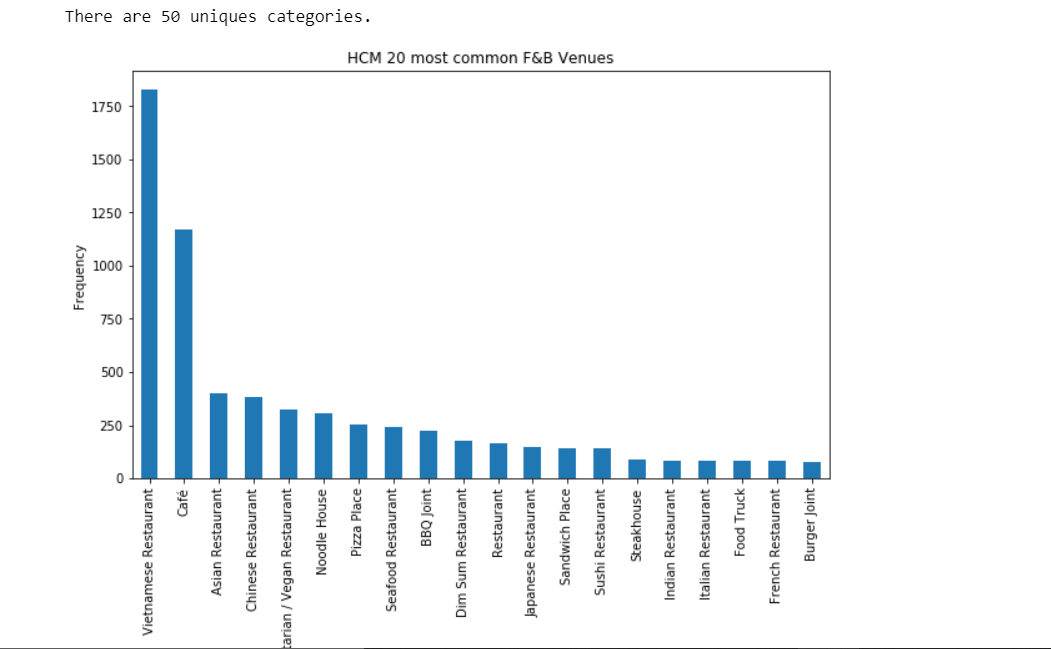
Get F&B venues in each ward using the “getNearbyVenues” function in the Lab sections. I made some changes as follow:

* I defined the radius of surrounding area as 5Km because it’s an average distance between each ward, to make sure that I don’t miss any venue.
* Remove the LIMIT as I’ll need all the venues.

The final venues table is as follow:

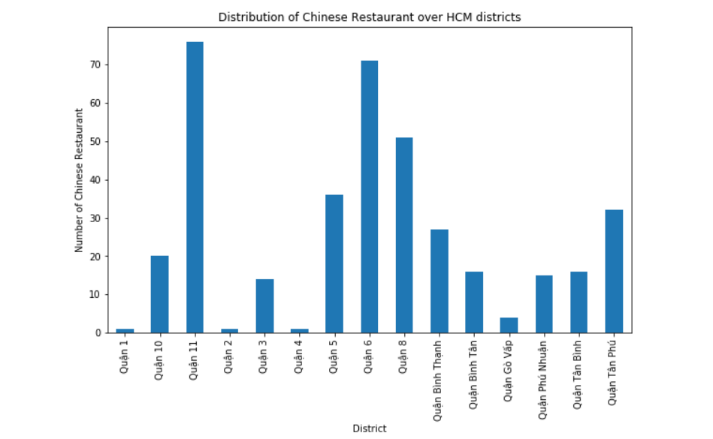
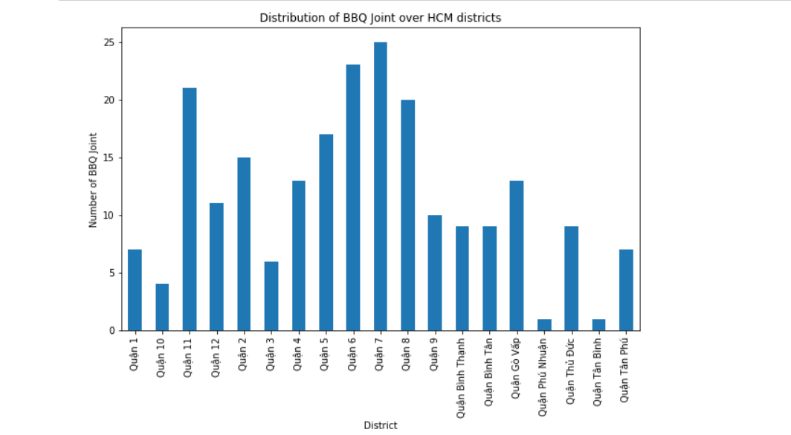


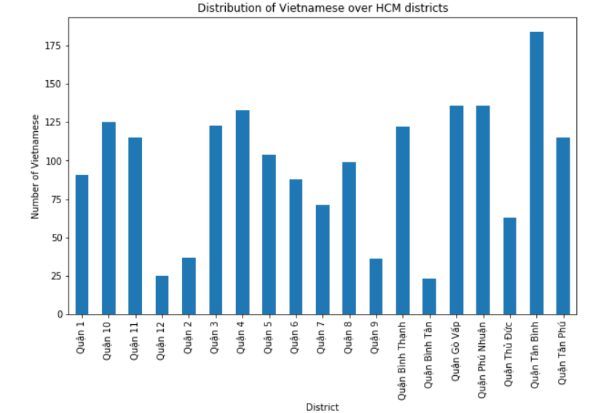
Let’s take a look at our data by finding out how many unique categories are there and frequency of each. I’ll plot the 20 most common F&B venues:



Vietnamese Restaurant and Café are the 2 most common in HCM city with more than 1750 and 1000 venues respectively, while other types of F&B have only around 250 venue counts.

Let’s take a look at how some of the venue types are distributed over HCM districts, namely Vietnamese Restaurant, Chinese Restaurant and BBQ Joint:

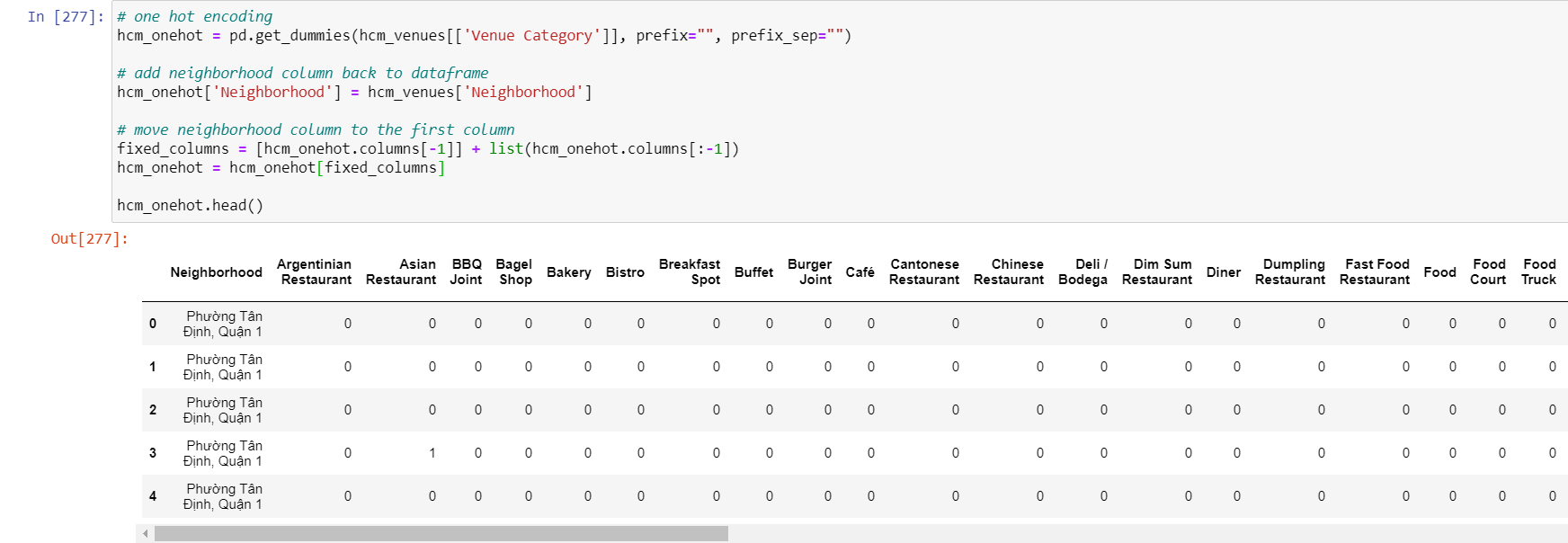




We can see that for each type of F&B venues there are certain district with relatively less venue count than others, which are the potential location for new restaurant with less competition. However, because we’re looking at venues in each district, which are the “clusters” where borders were not originally created for our venue purposes. Therefore, let’s work on clustering these venues using machine learning (in this case, k-means clustering) to get better results.

## Preprocessing data for K-means Clustering

Create one hot encoding mask



Calculate the means of frequency of occurrence of each venue type in each ward.



Create a data frame and display the top 10 venues for each wards

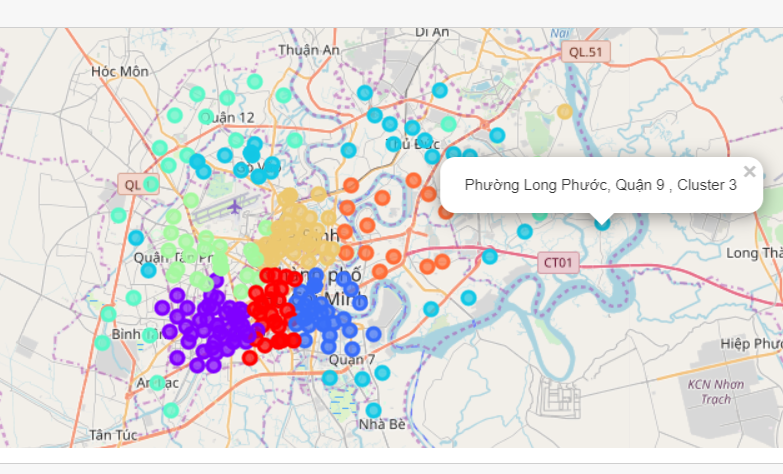


## K-means Clustering

I used this following code to choose the best K for K-means Clustering base on WCSS method

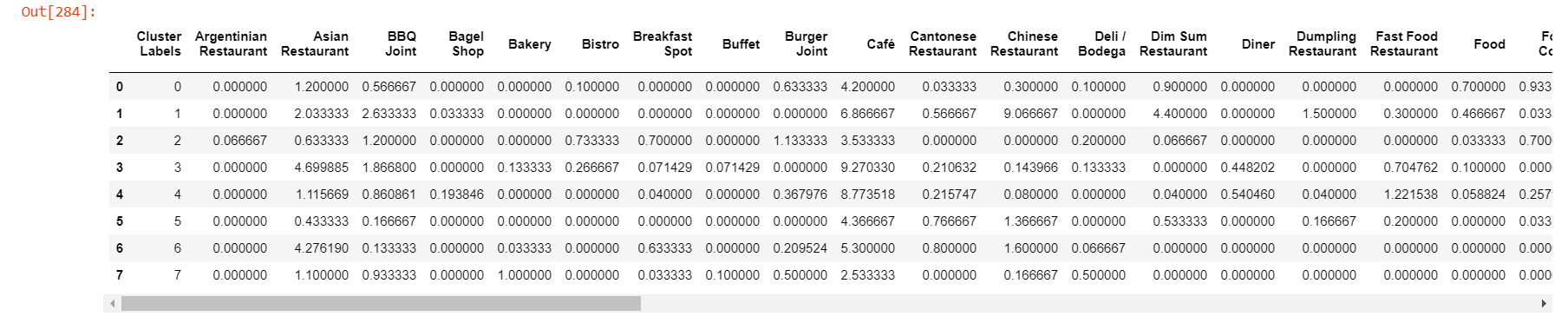


The graph indicates that the larger the K, the more accurate the clustering method is, thus I chose K>7. Next, cluster maps were plot for all K ranging from 1 to 10. Base on the cluster map as seen below (K=8), I chose K = 8 as the border between each cluster is the most well defined.



## Analyzing each cluster

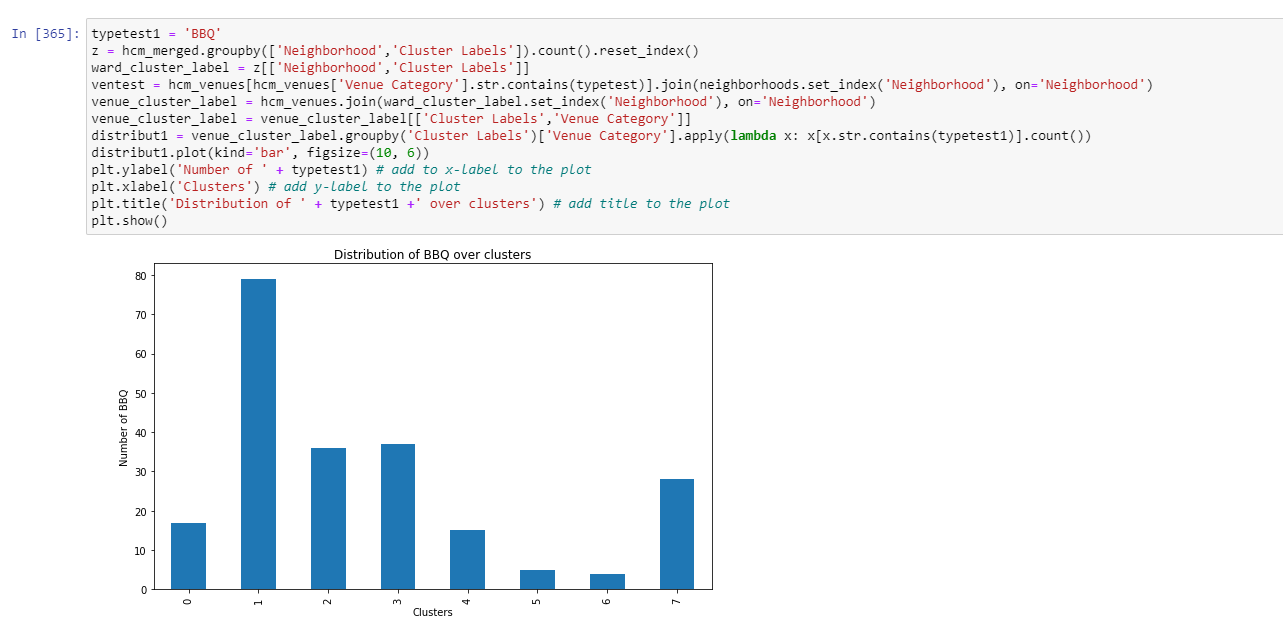
Cluster labels were added to each single venue as follow:



Base on that I created a table to sum up the 10 most common venues in each cluster:



From here we can explore the distribution/location of any type of F&B venue that we intended to invest in. For example, I’m going to open a new BBQ place and I’d like to look at the current BBQ restaurant distribution over all 8 clusters.



The graph show that cluster 5&6 have the least number of BBQ places. I’d prefer to open a new BBQ restaurant in these area for the least competition.

# Result & discussion

We’ve gone through step by step analysis of F&B venue location data in HCMC. The idea was to get the overall picture of F&B venues location & types and give recommendation to open new restaurant business. By using web scraping, geopy library and Foursquare API, I’ve found out the following:

* Most frequent venues of HCMC are café and Vietnamese restaurant, it’s easy to understand as Vietnamese cuisine & coffee culture is rich and diverse.
* Cluster 1 is only cluster with Chinese restaurant as most frequent venue instead of café or Vietnamese restaurant. This can be explained by the origin of people in the area of this cluster, which are mainly Chinese – Vietnamese. Thus, this area is basically the China Town of Ho Chi Minh city.
* Cluster 2: the center of Ho Chi Minh, where most foreigners live, which result in mostly international venues.
* Cluster 3 – the rural area of HCM city - which portrayed as a border around the main districts, tend to have similar popular food venues.
* Cluster 4,7: Café is the most frequent venue. Surprisingly in real-life these are the Cafe town of HCM city.

I’ve also made some comparison for specific type of restaurant among clusters. For example, the BBQ Joint venues are the most popular in cluster 1, I’d recommend avoiding this area if you’re going to open a BBQ restaurant. The method can be applied for other type of restaurant as well.

Also, because Foursquare is not popular in Vietnam, its data is very limited for city like Ho Chi Minh city. The location data required are insufficient and focus only on the main districts, which are frequently visited by foreigners/ Foursquare users.

Moreover, the category types are sometimes overlapping. For example: “Sushi restaurant” can also be under “Japanese restaurant”, which might result in miscalculation. I’ll work on grouping these venues together next time.

# Conclusion

Although the Foursquare data is still limited to a certain level, the location analysis gave me an overview on how F&B venues are distributed among HCM city. Beside venues segmentation, this analysis can also be used for insights on specific type of restaurants that are planned to be opened. Despite some drawbacks and room for improvement as mentioned, some of the result suit the real-life observation (for instance the China Town area with the most Chinese restaurants). I hope this will be valuable for all those who are interested in F&B venues, and as an example of applied data science in actual problems.