Finding Candidate Locations for Starting a French Restaurant in Hong Kong

Hyunmyung Myung

December 27, 2019

1. Introduction

This project assumes the following scenario. A group of stakeholders have consulted to find the best location for opening a new branch of French restaurant. Their initial small bistro in Paris has grown into a Michelin-starred, multinational business that spans across United States, Brazil, Germany, and other countries in the Western hemisphere, and now they want to expand their portfolio by opening up a new branch in the East. Hong Kong was chosen because the city is known for having the largest concentration of high-net-worth individuals of any city in the world[1], and its demographics seemed compatible with the business model of the stakeholders (3 E's they called it: exquisite, enamoring, and expensive). Hong Kong would be a real-world lab where they would experiment to see if the populations of the Eastern cities would be enamored to take the bite of 3 E's.

To find out the optimal location for opening a restaurant branch, three factors were considered:

(1) Proportion of French restaurants to all types of restaurants: this factor indicates the potential demand for French restaurants in each area -- lower proportion, higher likelihood. Proportion is used here instead of the sheer number of French restaurants, because less numbers may mean that there just are not enough demands due to less foot traffic, lower residential density, etc.

(2) Price points of the French restaurants in the vicinity: the business of the stakeholders aims for the best cuisine that money can buy, no moderation on ingredients or human resources whatsoever. This means the effect of a budget French restaurant on the potential demand of the stakeholder's restaurant would be less than a French restaurant with a similar price point.

(3) Ratings of the French restaurants in the vicinity: nearby French restaurants with lower ratings would have less impact on the profitability of the new branch to be opened by the stakeholders, and vice versa.

The above three points was be used as guidelines for finding optimal locations within Hong Kong. These variables shall be quantified based on the information extracted online using a platform that provides relevant insights.

2. Data

The area names are extracted from a Wikipedia page[1] and the coordinates are generated using Google Geocoder API for Python. All information related to venues is extracted using a personal account for Foursquare API. Other data that are not listed here have their sources attached to the location of its data placement.

3. Methods

3.1. Coordinates

The first data to be attained are the coordinates of various areas that fall under the districts of Hong Kong. The most comprehensive, though not complete, of the found data were on a Wikipedia page (https://en.wikipedia.org/wiki/List_of_places_in_Hong_Kong). This list should work out to provide coordinates for locating relevant venues via Foursquare. The following cell imports the page using Beautiful Soup in a Python-readable form. Next, the below cell lists a custom function that converts the above Beautiful Soup output to a data frame of area names.

Using the custom function as defined above, the coordinate values of each area were retrieved using the addresses generated by processing the information in the Wikipedia soup. These coordinates were used as geological points at which venue information were searched using Foursquare. Each instance of the custom function is written in a separate cell, because it takes time to retrieve the coordinates. It can be observed above that the geocoder module was unable to retrieve the coordinates for some of the address items. To deal with this limitation, the non-retrieved "not a number" entries were removed by dropping associated rows. Then, the retrieved coordinates are visualized using Folium.

The maps were generated, which showed that some coordinates are too close to one another to be valuable as separate locations for Foursquare searching. These overlapping coordinates were united into larger units using DBSCAN to avoid redundant analyses. After DBSCAN, the centroid of each cluster was calculated. Doing so revealed that the collection of coordinates seems less crowded than the pre-cluster version. Effectively, the number of coordinates has reduced to a manageable level, and the Foursquare API was employed to determine the viability of opening a French restaurant at each coordinate.

3.2. Foursquare API

Firstly, the proportion of the number of French restaurants to that of all types of restaurants was generated. Because harboring no restaurants in an area (especially in a hyper-developed state like Hong Kong) means there is probably no demand for restaurants in that area, the rows that have zero restaurants were dropped from the data frame.

---

[1] https://en.wikipedia.org/wiki/List_of_places_in_Hong_Kong

As could be seen from the shape of the resulting data frame, the number of candidate locations has been reduced to 59. However, it seems that the proportion cannot be used as meaningful data, because the number of venues that can be retrieved by each call is limited to 50. Therefore, some of the proportion values were 1.0. To bypass this limitation, the proportion variable was substituted by the number of French restaurants to approximate the density of the restaurants of this type.

Finally, Foursquare's premium calls were performed to collect additional information on the coordinates, which includes the average price point of the restaurants and the average rating of the French restaurants (to be collected in this order). The average price of the restaurants is collected using the "explore" function of the Foursquare API. To extract this data just using regular calls, the "price point" parameter of the explore call is used to count of the number of restaurants for each point and then to calculate average price point of each location. A box plot of the price points is rendered to show their distribution. It shows that most data lie between 1 and 2, with two outliers on the higher side.

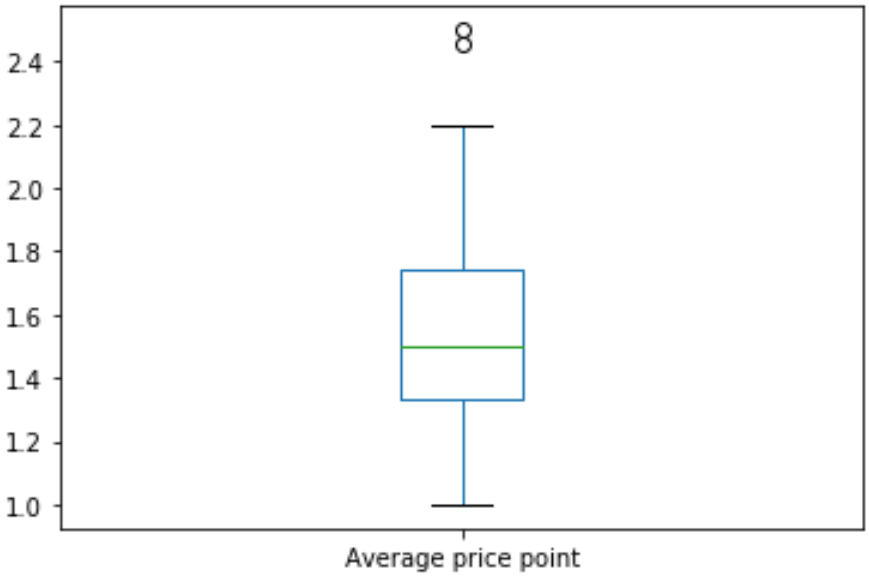


Figure 1. Box plot of the average price points

Next, the average rating of the French restaurants was calculated using the premium call "venue details" of the Foursquare API. Please note that, to conserve the number of premium calls, the retrieval limit has been intentionally lowered to 5. Admittedly, the accuracy of the data should naturally tend to increase with the increasing number of retrieval limit, and this should be the case if there is no limit to the number of calls.

3.3. K-means

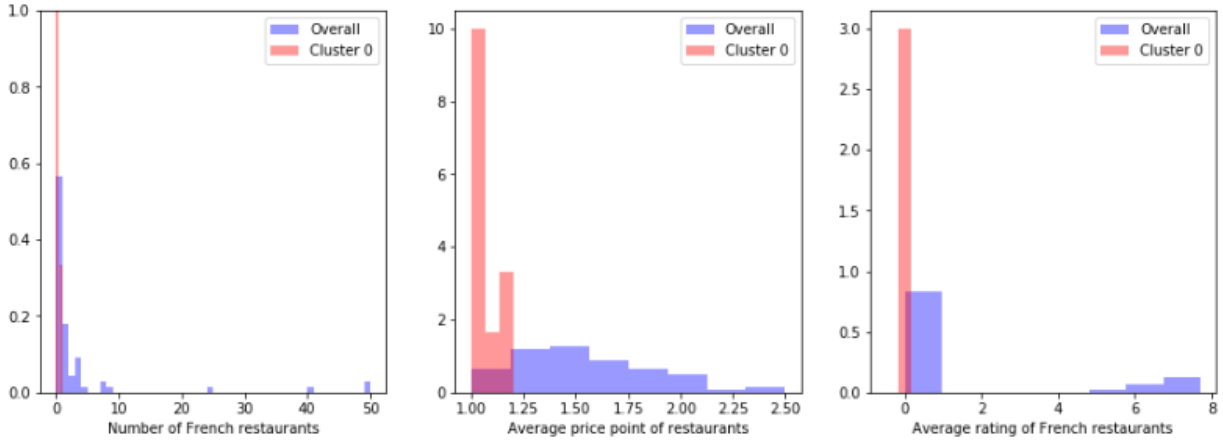
Now that all required data have been collected using Foursquare API, the data frame was processed to allow for clustering of the data points. The clustering method to be employed is K-means, to ensure that all the data points belong to one of the clusters.

4. Results

In this section, the characteristics of each cluster resulted from k-means are examined by observing the values of the variables that were defined and collected in previous sections. The variables, to reiterate, include: (1) the density of French restaurants (indicating the degree of interest overlap from similar venues), (2) the average price point of the restaurants (showing how affluent the customers in the areas are), and (3) the average rating of French restaurants (gauging the degree of the competition from similar venues). The statistics of each variable (listed in the table below) is used as the basis of comparison from which the other results are interpreted.

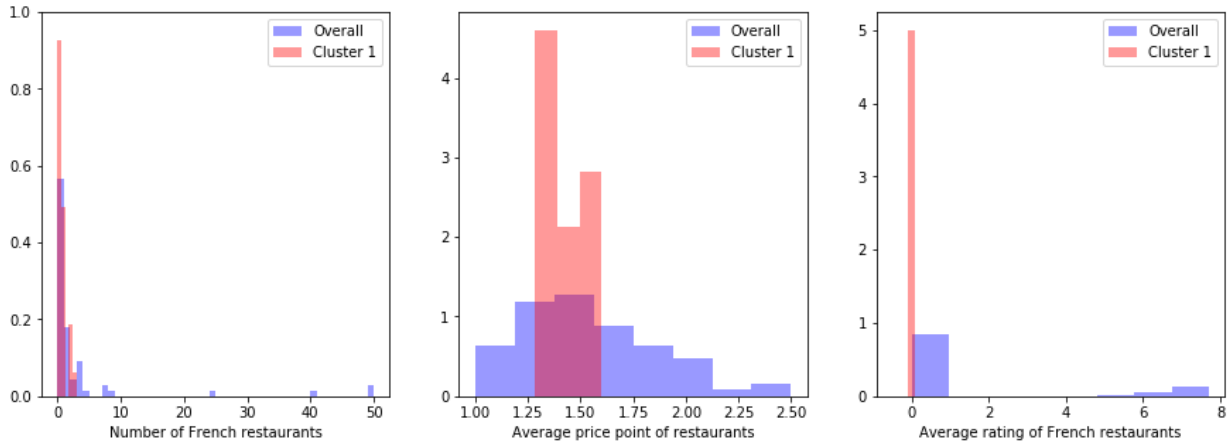
4.1. Cluster 0

The plots show that, in comparison to the overall trends in Hong Kong, the regions in this cluster have few to zero French restaurants (negligible because the average rating is all 0) and the nearby restaurants tend to have lower price points. This means that the neighbors in this cluster were not compatible with the high-price business model of the client, and thus these regions were not considered further.



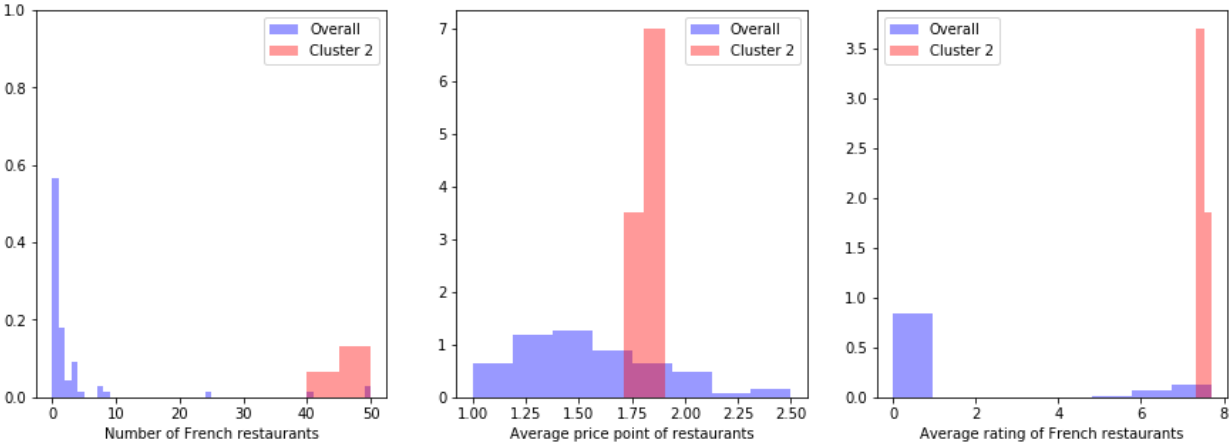
4.2. Cluster 1

The characteristics are highly like cluster 0 except that the price points are higher. Also were not considered further.



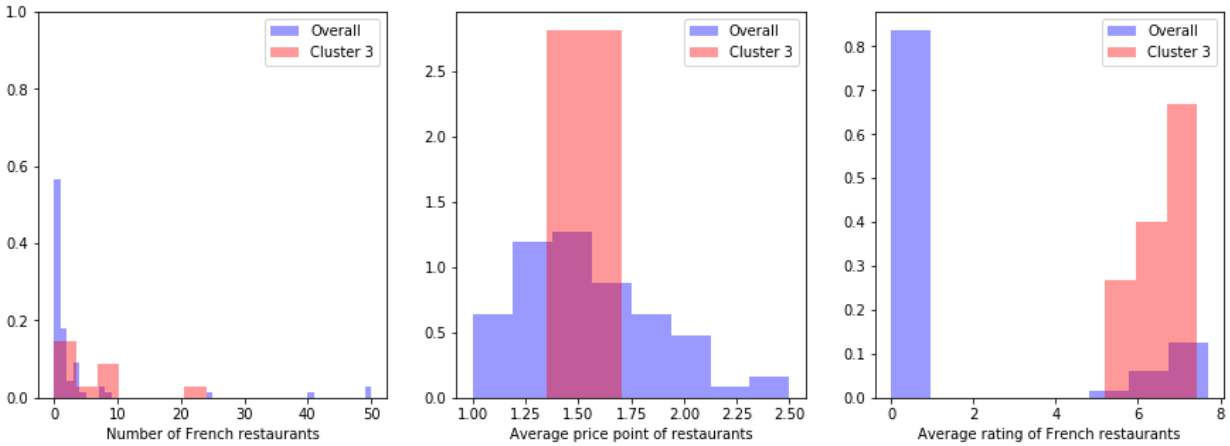
4.3. Cluster 2

This cluster is marked by the highest number of French restaurants, most of which very high ratings. In addition, the average price point of the restaurants is also higher than the other two clustered previous considered. The map reveals that this cluster consists of locations that are considered the busiest in Hong Kong, such as Sheung Wan, Wan Chai, etc.



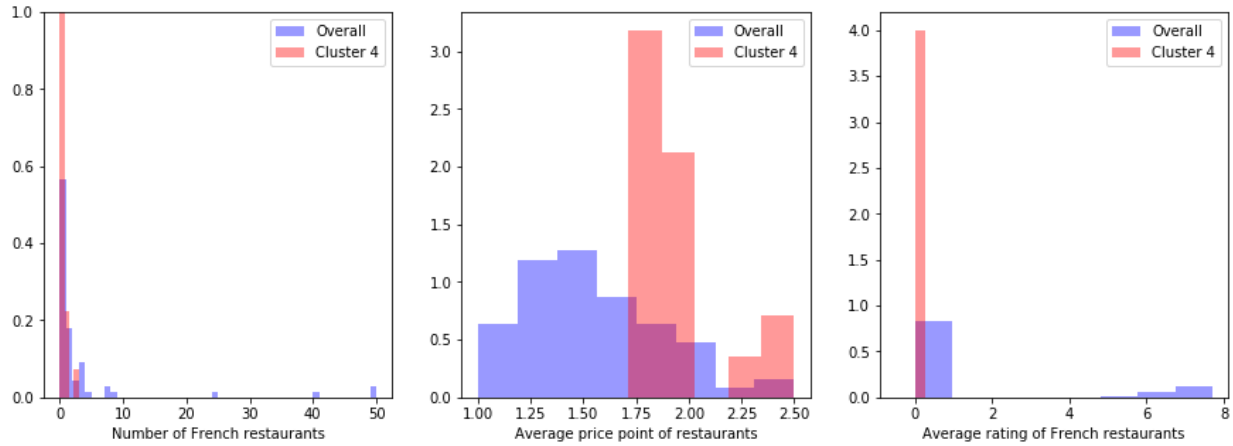
4.4. Cluster 3

This cluster has zero to some French restaurants with the average rating more diversified than cluster 2. The competition from other French venues imposed on the stakeholders were less than cluster 2, thanks to the varied distribution of the measured characteristics.



4.5. Cluster 4

The last cluster, unlike the others, are marked with few to zero French restaurants with the higher percentile of the price point of the nearby restaurants. The pre-established French restaurants have zero ratings.



5. Discussion

Before interpreting the results, the expectations of the stakeholders were recapped first. The French restaurant business run by the stakeholders offers expensive menus and target customers who are willing to experience the best of the Michelin-starred gastronomic experiences without worrying about price tags. In other words, the stakeholders would prefer to open their first branch in an area where there are enough people who can afford the menus. Hong Kong, however, already has other businesses that already offer similar experiences, firmly rooted in business-heavy areas marked by cluster 4. The stakeholders, knowing that their self-interest would be compromised by other existing French restaurant businesses, said they want to start their branch somewhere that would pose little to no competition.

With the above bolded statements in mind, it seems that the cluster that best fits the expectations of the stakeholders is the cluster 4. This cluster were further dissected to narrow the number of candidate locations to a handful. The below data frame of the cluster 4 entries shows that two of the variables (number of French restaurants and the average rating of French restaurants) do not vary enough to influence the decision-making process. Therefore, the average price point is the only one of the selected variables that can and should affect where the stakeholders want to open their branch. This said, the stakeholders were recommended to visit the places in the order listed in the data frame, from top to bottom. The below figure shows the top 3 locations that were recommended to the stakeholders.

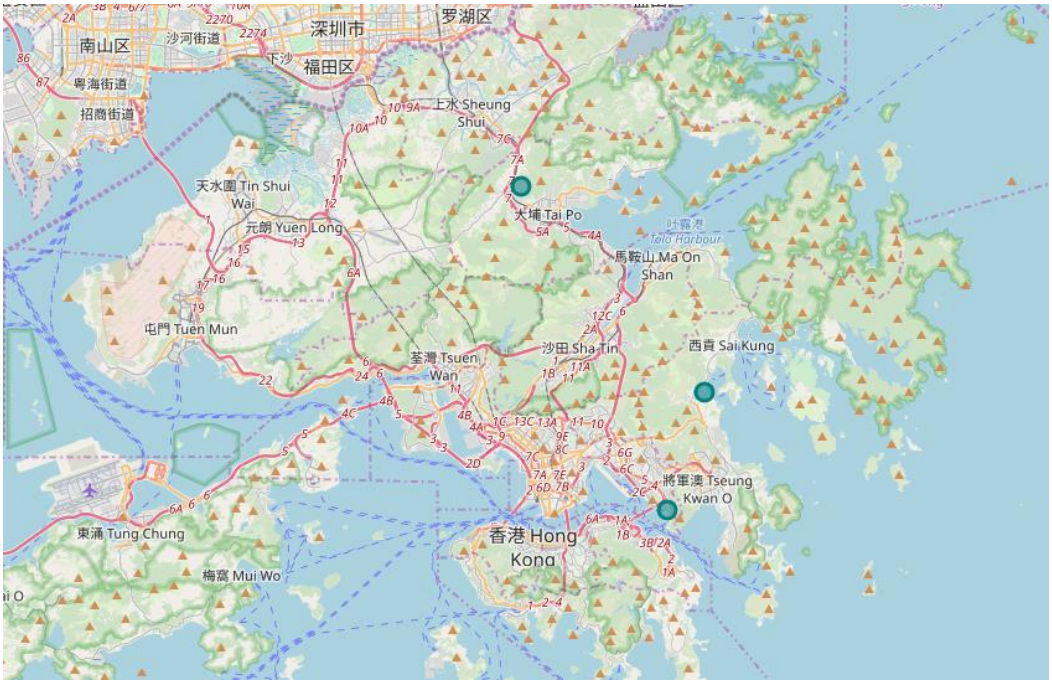


Figure 2. Top 3 location candidates for opening up the business of the stakeholders

6. Conclusion

The initial goal of finding the locations was successfully conducted by the end of the discussion section. It would better be noted that, although the goal was met, the means by which it was achieved could be optimized by streamlining a few processes that can be considered subpar. For example, the list of areas retrieved from the Wikipedia page, although the most comprehensive of the lists encountered, could be replaced simply by placing coordinates at constant intervals without any regard to the areas in Hong Kong. However, the advantage of using the Wikipedia page was that, because the region is filled with conserved natural territories despite being a highly metropolitan state as a whole, a lot of these coordinates of constant intervals would be rendered useless, because there would not be enough venue information to be usable for analysis.

Another point worthy of mentioning is that one of the variables that was pre-defined in the Introduction section (i.e. “Proportion of French restaurants to all types of restaurants”) was not obtainable with the method employed, because the Foursquare API limits the number of retrieved items to 50. Therefore, there was no means to knowing how many venues of a category exactly existed in an area if the number of these venues exceeded 50. In future endeavors, it is recommended that other methods be explored to bypass these limitations.

The number of calls that could be made was limited based on the calls that were capped by the Foursquare service provider. As the authentication used to retrieve the venue data was of a personal account type, the full range of the venue information was not retrieved. To further increase the accuracy of the analysis, it would be preferred if the cap on the number of calls can be lifted by creating an account that has more relaxed limits on the number of calls and other features.