Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Restaurant in HongKong

By: Rachel Gan

Octorber 28

1. Introduction

1.1 Background

For most people who want start a business, opening a resturant is a good choice, beacuse the catering industery is related to everyone and has a huge consumer base. Additional, the cost and risk of opening a resuturant is very low.
HongKong has a population of 7.5 million and a population density of 6,300 people per square kilometer, HongKong is also a multi-ethnic society, which means that the catering industering has large demands and diverse needs.
Before opening a restuarant, we need to take a lot of factors into consideratio to help us make a better decision. Particularity, the location of the restaurant is one of the most decisions that will determine wherther the reatuarant will be a success or a failure.

1.2 Bussiness Problem

The purpose of this top project is to analyze and select the best location in Hong Kong to open a restaurant. The project uses data science methods and machine learning techniques such as clusering to provide solutions to the business problem: If a entrepreneur wants to open a restaurant, where would you recommand the location is?

1.3 Interest

Entrepreneurs or companies that open restaurants may be interested in the location of restaurants, and other real estate agents may also be interested in.

2. Data

2.1 Required data

* List of neighborhoods in HongKong.
* Latitude and longitude coordinates of those neighborhoods.
* Venue data

2.2 Data sources

This wikipedia page (https://en.wikipedia.org/wiki/Category:Places_in_Hong_Kong) contains the list of neighborhoods in HongKong, with a total 67 neighborhoods. We will use Python requests and BeautifulSoup package to extract the neighborhoods data from the wiki page. Then we will

get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us the latitude and longitude of each neighborhood. After that, we are going to use Foursqauare API to get venue data of all neighborhoods, which will provide many categories of the venue data, we will focus on the restuarant categories in order to help us solve the futher problem.
Additional, we will also use many other data science skills to help us solve problems, such as we will use machine learning techniques like k-means clustering, data visualization packages like Folium, etc...

3. Methodology

3.1 Data sources

The first step is to use Python requests and BeautifulSoup packages to scrap the wikipedia we mentioned above, we can get a list of neighborhoods' names. Then we use Geocode pacakage to convert each neighborhood's address into geographical coordinates in the form of latitude and longitude. Now we get the neighborhoods' names, latitudes, longitudes information, we need to put the data into pandas DataFrame and the visualize the neighborhoods in a map using Folium package, according to the map, we can make sure wether we have found the correct coordinates by Geocode.

3.2 Data Collection

After that, we are going to use the Foursquare API to get top 100 venues within a radius of 2000 meters. We need to register a Foursquare developer account to get Foursquare ID and Foursquare SECRET. Then, we make an API call to Foursquare through the geographic area to call the coordinates of the neighborhood in the Python loop. Foursquare will return the venue data format in JSON, we will extract venue name, venue category, venue latitude and longitude. With the data, we can check how many venues each community has returned, and see from all the returned venues that many unique categories can be planned. Then, we will analyze the occurrence of each category of places by grouping rows by neighborhood and taking the mean frequency. In this way, we also prepared data for clustering. Since we are analyzing "restaurants" data, we filter "restaurants" as locations community category.

3.3 Data processing and analysis

Finally, we will perform clustering on the data using k-means clustering. The K-means clustering algorithm identifies k centroids, and then assigns each data point to the nearest cluster while keeping the centroid as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms, especially suitable for solving the problems of this project. According to the frequency of appearance of "Restaurant", we divide communities into three

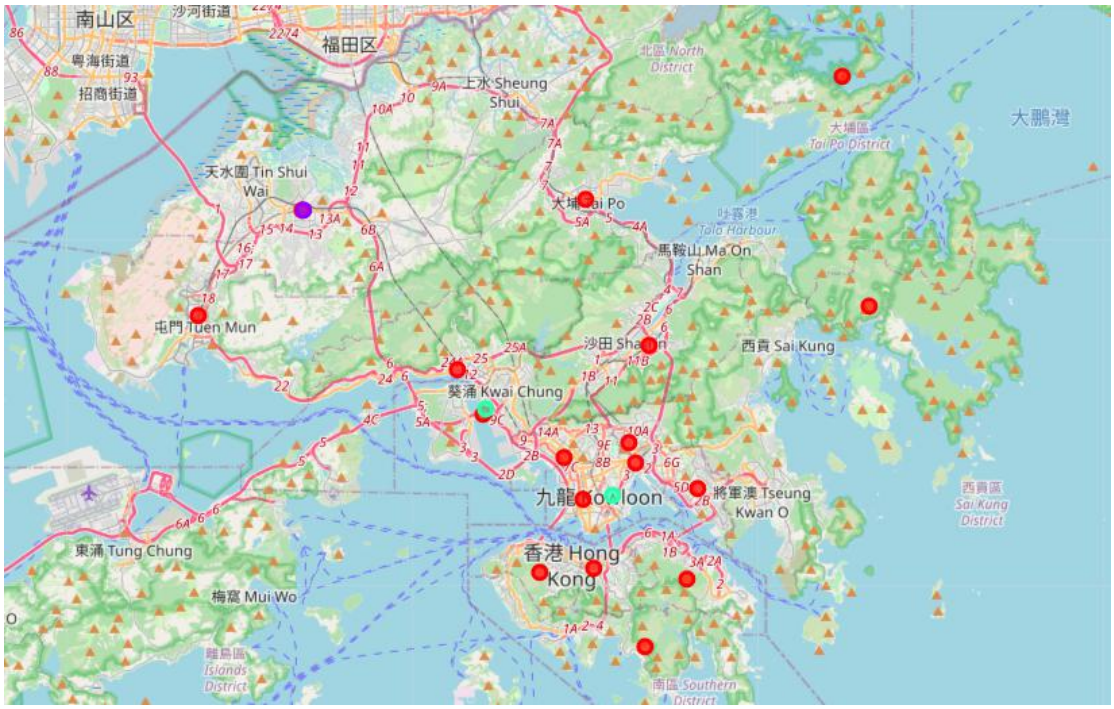
categories. The result will allow us to determine which neighborhoods have a high concentration of restaurant and which neighborhoods have fewer restaurant. Based on the emergence of restaurants in different communities, this will help us answer the question of which neighborhood is best for opening a new restaurant.

4. Conclusions

The result of k-means clustering shows that we can divide the neighborhood into 3
Clustering based on the frequency of occurrence of "restaurants":

- Category 0: A community with a concentrated number of restaurants
- Category 1: Neighborhoods with few or no restaurants
- Category 2: communities with few or no restaurants

The clustering results are visualized in the map below, cluster 0 is red, cluster 1 is
Purple, cluster 2 is mint green.



5. Discussion

From the observations observed on the map in the “Conclusions” section, most restaurants are concentrated in the downtown area of HongKong, with the largest number in cluster 0. On the other hand, categories 1 and 2 have very few nearby restaurants. This is an excellent opportunity to open a restaurant and a high-potential area because there is almost no competition in the existing restaurants. From another point of view, the conclusions also show that the oversupply of restaurants mainly occurs in the center of the city, while there are still very few restaurants in

the suburbs. Therefore, the project recommends that real estate developers make full use of these findings and open new restaurants in the communities of clusters 1 and 2 with little or no competition. It is recommended that real estate developers avoid communities in cluster 0 that already have a large number of restaurants and are highly competitive.

6. Limitations and recommendations for future research

In this project, we only consider one factor, that is, the frequency of restaurant appearance. There are other factors, such as population and residents' income, which may affect the location decision of the new restaurant. However, as far as the researcher knows, such data cannot be used at the community level required for the project. Future research may devise a method to estimate such data that will be used in the clustering algorithm to determine the preferred location for opening a new restaurant.