IBM Data Science Professional Certification on Coursera

Capstone Project - The Battle of Neighborhoods

Finding Optimal Locations to Open an Indian Restaurant in Singapore

Final Report

Prepared by: Akshaya Suresh

Introduction/Business Problem

Singapore is one of the most diverse countries in the world. It is a small yet mighty city-state, that's home to a wide range of cultures, ethnicities and religions. This is especially evident in the wide availability of its cuisines, which predominantly come from the Chinese, Malay and Indian communities. This makes Singapore a very attractive hub for restaurateurs who are contemplating to open a restaurant that serves one of these cuisines. In this project, we aim to find the most optimal location to recommend to stakeholders who are planning to open an Indian restaurant in Singapore.

There are many things to consider when choosing a location, but some basic principles can help restaurateurs get a better understanding of what it takes to make up a good restaurant location. We will define an optimal location based on the following criteria:

Competition - This factor is a double-edged sword; being close to established competition may help with business marketing, but if the new restaurant is too close to its competition, it may have a tough time gaining a foothold in the community. Hence, we prefer neighborhoods that don't already have many Indian restaurants but at the same time don't have too few restaurants.

Accessibility - Another important factor is how accessible the potential location is. When looking at a restaurant location, we need to consider the amount of accessibility to make it as easy as possible for customers to visit the establishment. We should also keep in mind that tourists tend to visit eateries that are in or around the city center. Hence, we prefer more centrally located neighborhoods.

Data

To obtain the list of neighborhoods and their corresponding regions in Singapore, there exists a Wikipedia page titled “Planning Areas of Singapore” that has all the information we need to explore and cluster the neighborhoods in Singapore. We will scrape this Wikipedia page, wrangle the data and clean it to prepare it for use.

We will then proceed to use the Geocoder Python package to obtain the latitude and the longitude coordinates of each neighborhood.

Lastly, we will use the Foursquare API to explore the neighborhoods and segment them. We will analyze the venue data, cluster the neighborhoods with respect to the number of Indian restaurants, and finally examine the results to make recommendations based on the data.

Methodology

Singapore has 55 neighborhoods, organised into 5 divisions. This information is available in the form of a table in this Wikipedia page: https://en.wikipedia.org/wiki/Planning_Areas_of_Singapore. We scrape this Wikipedia page using Beautiful Soup, and then read the data onto a pandas dataframe so that it is in a structured format. We then remove the columns that we aren’t interested in, leaving only the neighborhoods and their corresponding regions left in our pandas dataframe.

Now, we use the Geocoder Python package to get the geographical coordinates of each neighborhood in our list. In order to make sure that we get the coordinates for all of our neighborhoods, we run a while loop for each neighborhood. Once we've gotten all the coordinates, we read this data onto a new pandas dataframe. We then merge both dataframes. Our resulting dataframe has the name of each neighborhood, its corresponding region as well as the neighborhood's latitude and longitude.

Next, we utilize Foursquare API to explore the neighborhoods. We first get the top 100 venues that are in each neighborhood within a radius of 1000 meters. We create a new dataframe for this venue data. To get a better idea of what our dataframe looks like, we check how many venues were returned for each neighborhood, and check how many unique categories can be curated from all the returned venues.

We then use one hot encoding to convert the categorical venue data to integer data. After which, we group the rows by neighborhood and by taking the mean of the frequency of occurrence of each category. We can finally create a dataframe to view only the mean of the frequency of Indian restaurants in each neighborhood, as this is our category of interest.

We use k-means clustering to cluster the neighborhoods in Singapore into three clusters. We create a new dataframe that includes the neighborhoods, regions, coordinates, cluster labels and mean of frequency of Indian restaurants. We can visualize the resulting clusters by creating a map of Singapore with the various clusters superimposed on top, using the Folium package.

Results

We will partition the neighborhoods into three groups since we specified the algorithm to generate three clusters. The neighborhoods in each cluster are similar to one another in terms of their means of frequency of Indian restaurants.

The three clusters are:

Cluster 0 - Neighborhoods with a low number of Indian restaurants
Cluster 1 - Neighborhoods with a high number of Indian restaurants
Cluster 2 - Neighborhoods with a medium number of Indian restaurants

Discussion

Competition

Cluster 1 only consists of Rochor, which is the neighborhood with the highest frequency of Indian restaurants in our dataset. This makes sense as Little India, a buzzing historic area that shows off the best of Singapore's Indian community, is situated in Rochor. It is a popular destination for tourists and locals alike to bask in its rich culture. So it comes as no surprise that Rochor has a significantly higher number of Indian restaurants as compared to any of the other neighborhoods. As such, this cluster is likely already suffering from high competition and would not be a good choice for restaurateurs to pick as a location for a new Indian restaurant.

Cluster 0 consists of neighborhoods that have few to no Indian restaurants. These neighborhoods may be considered by stakeholders, but more research should be done in order to examine these neighborhoods further; there might be a range of reasons as to why they have been unpopular venues for Indian restaurants.

Cluster 2 consists of neighborhoods with a moderate number of Indian restaurants. This is an indication that these neighborhoods have been viable options for restaurateurs to conduct their business in. But of course, similar to the neighborhoods in cluster 0, these recommended neighborhoods should be considered only as a starting point for more detailed analysis which will eventually result in an optimal location that has taken all factors into account.

Accessibility

Since we prefer more centrally located neighborhoods in order to cater to tourists, we give preference to the neighborhoods in cluster 2 (which has the most ideal amount of competition as discussed above) that are situated in the central location. These neighborhoods are Bukit Merah, Geylang, Marine Parade, Newton and Queenstown.

Conclusion

A limitation here is that all 55 neighborhoods in our dataset are of different areas. Some neighborhoods such as Museum, Orchard, Singapore River and Straits View, are each less than 1 square kilometer in area. However, neighborhoods such as Bedok, Queenstown, Tampines and Yishun are each more than 20 square kilometers in area. As such, using a 1000 meter radius to gather venue data from each of the 55 neighborhoods is problematic as they are all of unequal areas. This could be mitigated by creating a hexagonal grid of area candidates, equally spaced and centered around the central region of Singapore (instead of using the list of neighborhoods from Wikipedia as our area candidates).

The purpose of this project was to identify neighborhoods in Singapore that currently have a relatively low number of Indian restaurants. We first gathered the geographical coordinates of each neighborhood using the Geocoder Python package, calculated the Indian restaurant density distribution in each neighborhood using Foursquare API, and then clustered these locations in order to identify potential neighborhoods to recommend for our purpose.

The final decision on an optimal location for a new Indian restaurant should be made by stakeholders based on specific characteristics of every recommended neighborhood, taking into consideration additional factors like the demographics of the neighborhood, proximity to parking lots, surrounding traffic patterns, real estate availability and price.