Coursera Capstone

IBM Applied Data Science Capstone

Opening a coffee shop in Hong Kong, China, Asia

Apr2019

Introduction

Coffee is a brewed drink prepared from roasted coffee beans, the seeds of berries from certain *Coffea* species. The culture of having coffee have a significately increase in Hong Kong, China, Asia, in the past ten years, and this becomes a super sweet market for inventive Hong Kong entrepreneurs.

However, in what way the entrepreneurs could to reach the maximum profile for their coffee shop investment? Location to run the business is definitely one of importance factor.

Business Problem

The objective of this capstone project is to analyze the best location to start up a Coffee shop in Hong Kong, China. Using data science methodology and clustering by machine learning, this project would provide a suggestion to answer the business question: Where would be the recommended location to invest a coffee shop in Hong Kong, China.

Target Audience of this project

This project is particularly useful to anyone who would invest a coffee shop or plan to expend a coffee shop in city Hong Kong, where having a coffee craze. This project is timely as the city currently suffering from oversupply of coffee.

Data

In this section, the data need to provide solution are listed:

List of neighborhoods in Hong Kong. This defines the scope of this project which is confined to the city of Hong Kong, an international market of China in Asia.
Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.
Venue data, particularly data related to coffee shop. This is the data for clustering on the neighborhoods.

Sources of data and methods to collect

The list of neighborhoods in Hong Kong is sourcing from a Wikipedia page (https://en.wikipedia.org/wiki/Districts_of_Hong_Kong). Hong Kong with a total of 18 neighborhoods. Web scrap techniques is used to collect the data from the Wikipedia page under the help of Python requests and beautifulsoup packages.

Next is getting the geographical coordinates of the neighborhoods using Python Geocoder package which will return the latitude and longitude coordinates of the neighborhoods.

Last, the venue data of coffee shop will be sourcing with Foursquare API. Foursquare has one of the largest databases of 105+ million places and is used by over 125,000 developers. Foursquare API provide many categories of venue data, but in order to address our key question and align of this project scope, Boba shop is the key interest.

In summarize, this is a project make use of data science skills for analysis, including:

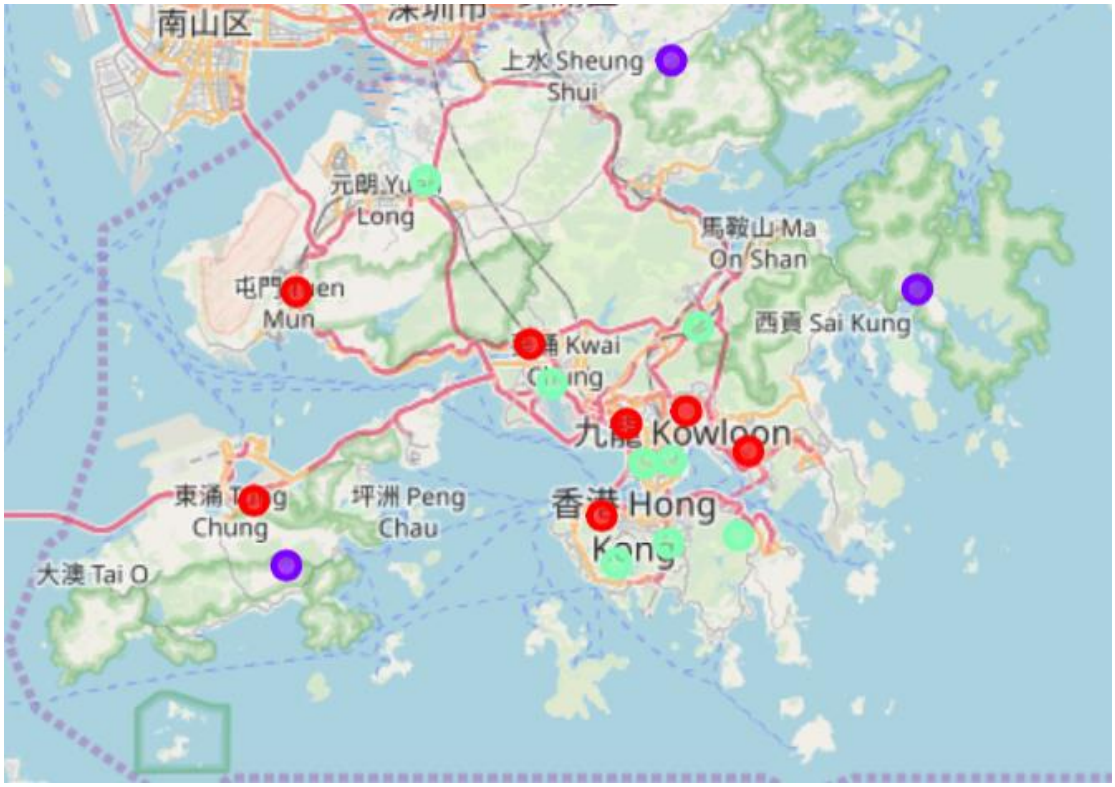
- Web scraping from Wikipedia
- API connection with Foursquare
- Data cleaning
- Data wrangling
- Machine learning on K-means clustering
- Map visualization with Folium

Results

The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for “Coffee Shop”:

Cluster 0 : Neighborhoods with high concentration of coffee shop
Cluster 1 : Neighborhoods with low number to no existence of coffee shop
Cluster 2 : Neighborhoods with moderate number of coffee shop

The results of the clustering are visualized in the map below with cluster 0 in red colour, cluster 1 in purple colour, and cluster 2 in mint green colour.



Discussion

Most of the coffee shop are concentrated in the central area of Hong Kong, with the highest number in cluster 0 and moderate number in cluster 2. On the other hand, cluster 1 has very low number to totally no coffee shop in the neighborhoods. This represents a great opportunity and high potential areas to open new coffee shop as there is very little to no competition from existing. Therefore, this project recommends enterprise to capitalize on these findings to open new coffee shop in neighborhoods in cluster 1 with little to no competition. Investor with unique selling propositions to stand out from the competition can also open new coffee shop in neighborhoods in cluster 0 with moderate competition. Lastly, it is advised to avoid neighborhoods in cluster 0 which already have high concentration of coffee shop and suffering from intense competition.

Limitations and Suggestions for Future Research

In this capstone, the question of where is the best location for a coffee shop opening would like to answer. In our observation, only consider one factor i.e. frequency of occurrence of coffee shop, but there should be others like population and rental payment etc. However, to the best knowledge of this analysis such data are not available to the neighborhood level required by this capstone. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new coffee shop.

Conclusion

In this capstone, the process of identifying business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders. To answer the business question that was raised in the introduction section, the answer proposed by this capstone is : The neighborhoods in cluster 1 would be a preferred location to open a new coffee shop with the lowest number of competitors. This findings will help the relevant start up to capitalize on the opportunities on high potential locations while avoiding overcrowded area in the decisions top open a new coffee shop.