Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Coffee shop around Universities in Beijing

By: Ruiqi Su

July 14th 2020

Introduction and Business Problem

Although Coffee is not that main beverage in China, but for modern society, Coffee shops is still an important kind of fast-retailing in big city. My undergraduate studying was in China University of Mining and Technology, Beijing. Most of Coffee shops here are Starbucks, Costa, and they are too expensive for our students. So this project is assuming that I am going to invest a Coffee shop or Chain brand Coffee for students in this area.

The objective of this capstone project is to analyse and select the best locations in the city of Beijng near Universities to open a new coffee shop. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In Beijing and region near Universities, if I am looking to open a new coffee, where is the best place?

Data

To solve the problem, we will need the following data:

- **List of Universities in Beijing.**

This defines the scope of this project which is confined to the city of Beijing, the capital of P.R.China.

- **Latitude and longitude coordinates of those Universities.**

This is required in order to plot the map and also to get the venue data.

- **Venue data, particularly data related to Coffee shops.**

We will use this data to perform clustering on the neighbourhoods.

Methodology

Firstly, we need to get the list of Universities in the city of Beijing, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/List_of_universities_and_colleges_in_Beijing).

We will do web scraping using Python requests and beautifulsoup packages to extract the list of universitiesdata. However, this is just a list of names. We need to get the geographical

coordinates in the form of latitude and longitude in order to be able to use Gaode API(Since some Chinese geo-info are not existed in Fouquare API. Gaode API is a Chinese Version of Google map API0. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the Universities in a map using Folium package. (We only use Gaode to get the coordinates, in venues part, we still use Foursquare API)

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the Universities in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category.

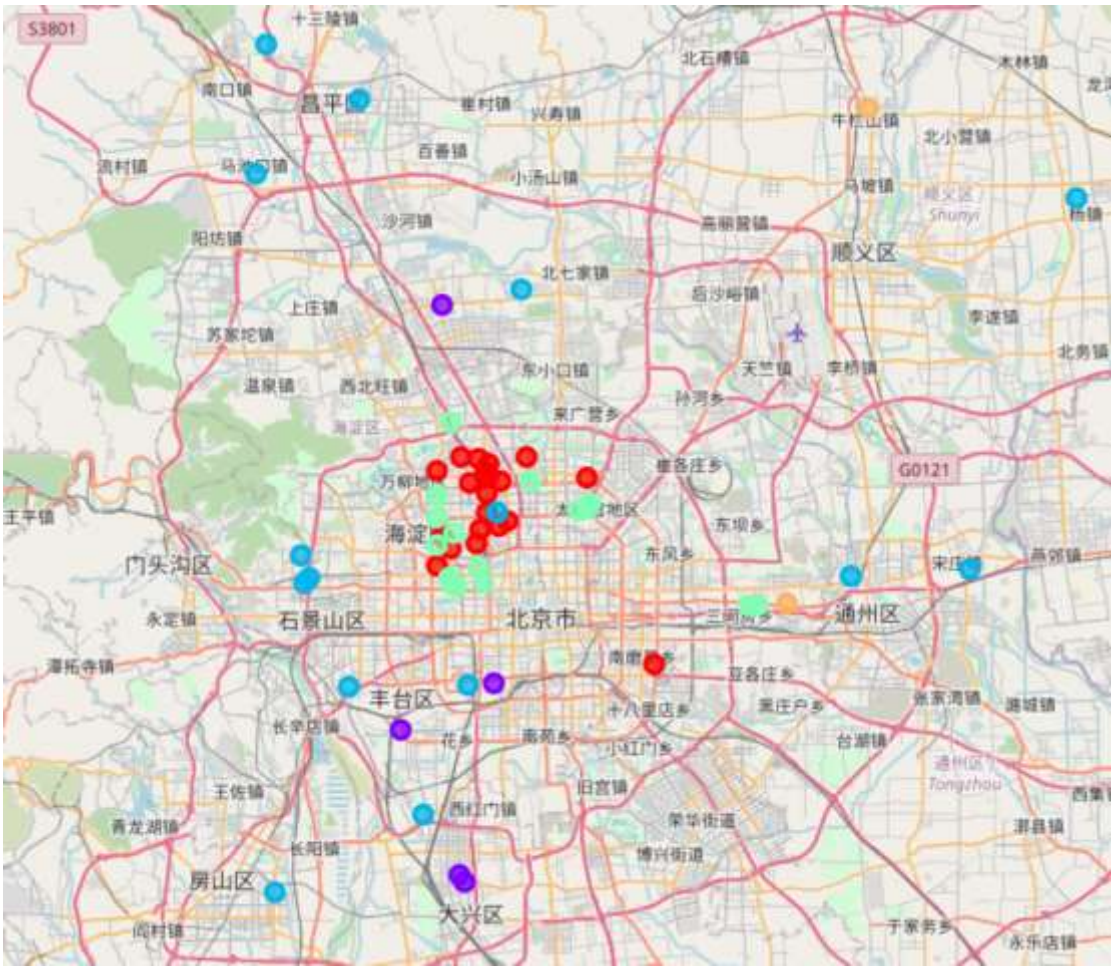
By doing so, we are also preparing the data for use in clustering. Since we are analysing the "Coffee Shop" data, we will filter the "Coffee Shop" as venue category for the neighbourhoods. Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 5 clusters based on their frequency of occurrence for "Coffee Shop". The results will allow us to identify which neighbourhoods have higher concentration of Coffee Shop while which neighbourhoods have fewer number of shopping malls. Based on the occurrence of Coffee Shop in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new Coffee Shops

Results

The results from the k-means clustering show that we can categorize the neighbourhoods into 5. Clusters based on the frequency of occurrence for "Coffee Shop":

- Cluster 0: Neighbourhoods with moderate number of Coffee Shop
- Cluster 1: Neighbourhoods with relatively high concentration of Coffee Shop
- Cluster 2: Neighbourhoods with low number to no existence of Coffee Shop
- Cluster 3: Neighbourhoods with moderate concentration of Coffee Shop
- Cluster 4: Neighbourhoods with high concentration 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, cluster 3 in mint green colour, cluster 2 in blue colour and cluster 4 in orange.



Discussion

As we can see, because most of Universities in Beijng are located in Haidian District, a large numbers of coffee shops are located in this region, with with moderate number in cluster 0 and cluster 2. But does it mean we have a great opportunity to open new coffee shops in this area? Not really. Coffee shops here are concentratede in office buildings and shopping malls, that are not too far from univeristy. Brands like Starbucks,Costa,Pcific Coffee have craved up the market, but mainly focused on white-collar not too much on students. Since students in University cannot pay that expensive coffee like 30-40 yuan per cup, so there are still potential for this sub-market.Luckin coffee has been always focused on it, but since Covid-19 and there are no students in campus,so a lot of School Luckin Coffee shops have been closed. However, there is still a large investment opportunity in this area for opening studentent Coffee shop.

Outside Haidian District, there are still several university cluster but they are relatively remote and they have two caes as below. First is remote but in center of that area;Second is remote but not in center of that area. In First case, like purpple dot(Cluster1) and orange dot(Cluster 4), investors should use the same strategy as in Haidian District. In second case, you should not focus on students but also other groups like withe-collars etc.

Limitations and Suggestions for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of Coffee Shop, there are other factors such as population and income of residents that could influence the location decision of a new Coffee shope. However, to the best knowledge of this researcher such data are not available to the neighbourhood level required by this project. 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 shopping mall. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 5 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new Coffee Shop. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 1 are the most preferred locations to open a new Coffee Shop, but must focus on the students' demands. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new Coffee shop.