Opening a New Shopping Mall in Beijing, China

Aixi Xu

September 3,2019

1. Introduction

Shopping malls are a great place for people to relax. People can spend a wonderful weekend with their families, where people can buy some essential necessities and some other useful things, so the mall must be set up in a crowded community. . But there are also other issues to consider, such as land prices, competition between malls, and so on. In addition to these complex factors, the final important factor in determining whether a mall can be profitable is how to choose a location. If we want to open a new mall in Beijing, the capital of China, how to choose the best location. Now my friend LEE wants to open a new shopping mall in Beijing. Below I will use data analysis to help him determine the best place to open a mall.

2. Business Problem

In order for Lee's shopping mall to make money, we need some information about Beijing. We must ensure that there are enough consumers in the mall to open, and there is no mall nearby to compete with him, so that he can earn a lot of money.

3. Data

3.1 Synopsis

The cities we will be analyzing in this project are: Beijing, China.
We will be using the below datasets for analyzing Beijing.

- The Localities of Beijing, China from Wikipedia: https://en.wikipedia.org/wiki/Category:Neighbourhoods_of_Beijing
- The coordinates (latitude, longitude) of these Localities of Beijing, OR from Open Street Map APIs
- From Foursquare we will need following venues data:
  - The shopping mall venues of the Localities.

We will then leverage the data in order to determine which locality is the most appropriate in order to locate the shopping mall.

3.2 Sources of data

This Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighbourhoods_of_Beijing) contains a list of neighborhoods in Beijing, with a total of 50 neighborhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us the latitude and

longitude coordinates of the neighborhoods. After that, we will use Foursquare API to get the venue data for those neighborhoods. Foursquare API will provide many categories of the venue data, we are particularly interested in the Shopping Mall category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

4. Methodology:

4.1 Python Libraries Used:

1. Pandas and Numpy for Data Analysis
2. Beautifulsoup for extracting data from Wikipedia Page
3. Geopy to extract Geo Coordinates from an address
4. Sklearn to use K-Means Clustering
5. Matplotlib and Folium for Data Visualization

4.2 Data Collection:

1. Go to the Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighbourhoods_of_Beijing).We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighborhoods data.
2. Use the Geocoder library to convert the neighborhoods obtained above into geographic coordinates in latitude and longitude.

Beijing Neighborhoods look like below:

3. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighborhoods in a map using Folium package.
4. We will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters.
5. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude.
6. Then, we will Analysis each neighborhood by grouping.

4.3 Data Analysis:

1. We are going to cluster and analyze the data using k-means clustering.
2. We are analyzing the "shopping mall" data and we will filter the "shopping mall" as the venue category for the community.
3. Clustering was performed using the k-means class algorithm. This is one of the convenient and popular unsupervised machine learning algorithms that are especially suitable for this project.
4. The community is divided into 4 clusters based on the frequency of the "shopping mall".
5. Select a region with the most suitable mall density to open a new shopping mall (at least as a candidate).

5. Results

5.1 K-Means Clustering

A K-Means clustering with 4 clusters have been performed

Here is the data look like

Cluster 3

	Neighborhood	Shopping Mall	Cluster Labels	Latitude	Longitude
6	Chuiyangliu	0.048387	3	39.88878	116.46472
15	Fuchengmen	0.045455	3	39.92295	116.34780
25	Jianguomen	0.040000	3	39.91460	116.41671
4	Chaoyangmen	0.040000	3	39.91460	116.41671
22	Hepingmen	0.033708	3	39.89996	116.37435
14	Dongzhimen	0.040000	3	39.93596	116.43027
7	Dashanzi	0.040000	3	39.98635	116.48395
16	Fuxingmen	0.056338	3	39.90556	116.35111

Cluster 2

	Neighborhood	Shopping Mall	Cluster Labels	Latitude	Longitude
41	Yayuncun Subdistrict	0.0	2	40.01388	116.39644
1	Andingmen	0.0	2	39.94382	116.39952
37	Xinjiangcun	0.0	2	23.20555	113.53199
35	Wudaokou	0.0	2	39.99257	116.33208
34	Weigongcun, Beijing	0.0	2	39.95736	116.31273
32	Tiantongyuan	0.0	2	26.78194	112.13472
21	Hepingli Subdistrict, Beijing	0.0	2	39.95276	116.41093
10	Deshengmen	0.0	2	39.94126	116.37929
11	Di'anmen	0.0	2	39.93943	116.39301
42	Yongdingmen	0.0	2	39.87028	116.39306
19	Guang'anmen	0.0	2	39.88806	116.34194
17	Gaoliying, Beijing	0.0	2	40.17057	116.49630

Cluster 1

	Neighborhood	Shopping Mall	Cluster Labels	Latitude	Longitude
24	Huilongguan	0.103448	1	40.07718	116.33527

Cluster 0

	Neighborhood	Shopping Mall	Cluster Labels	Latitude	Longitude
0	798 Art Zone	0.020000	0	39.90750	116.39723
40	Yabaolu	0.020000	0	39.90750	116.39723
39	Xuanwumen (Beijing)	0.020000	0	39.90750	116.39723
38	Xizhimen	0.018519	0	39.93889	116.35028
36	Xidan	0.020000	0	39.90750	116.39723
33	Wangjing Subdistrict	0.030000	0	39.99330	116.47284
31	Shifoying	0.020000	0	39.90750	116.39723
30	Sanlitun	0.020000	0	39.93609	116.44375
29	Ping'anli	0.020000	0	39.90750	116.39723
28	Niujie	0.027027	0	39.88301	116.35703
27	Neighborhoods in Beijing	0.020000	0	39.90750	116.39723
26	Madian, Beijing	0.020000	0	39.90750	116.39723
23	Huashi, Beijing	0.020000	0	39.90750	116.39723
20	Guomao, Beijing	0.020000	0	39.90750	116.39723

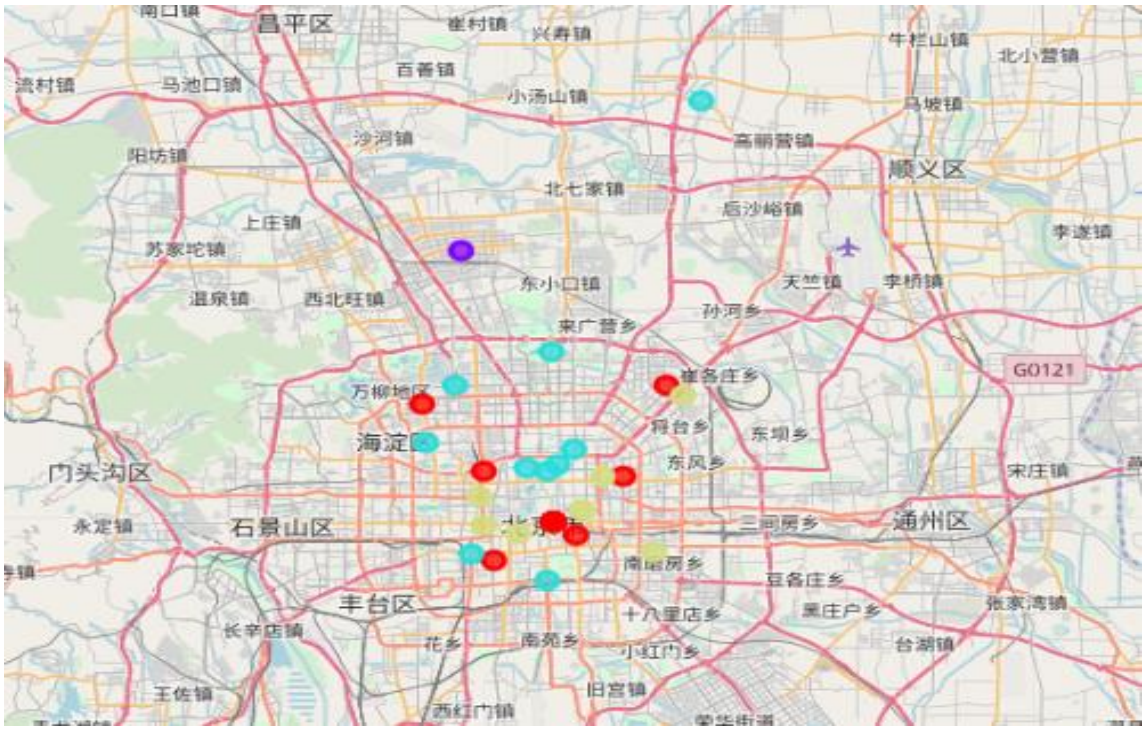
18	Gongzhufen	0.020000	0	39.90750	116.39723
43	Zhongguancun	0.030000	0	39.98111	116.30889
8	Dashilan Subdistrict	0.020000	0	39.90750	116.39723
2	Beijing central business district	0.020000	0	39.90750	116.39723
13	Dongsi Subdistrict, Beijing	0.020000	0	39.90750	116.39723
3	Brown Stone	0.020000	0	39.90750	116.39723
12	Dongdan, Beijing	0.020000	0	39.90750	116.39723
5	Chongwenmen	0.030000	0	39.89972	116.41222
9	Dengshikou	0.020000	0	39.90750	116.39723

5.2 Data Visualization

The results from the k-means clustering show that we can categorize the neighborhoods into 4 clusters based on the frequency of occurrence for “Shopping Mall”:

- Cluster 0: Neighborhoods with moderate number of shopping malls
- Cluster 1: Neighborhoods with high concentration of shopping malls
- Cluster 2: Neighborhoods with low number of shopping malls
- Cluster 3: Neighborhoods with moderate number of shopping malls

The results of the clustering are visualized in the map below with cluster 0 in red colour, cluster 1 in purple colour, cluster 2 in yellow-green colour, cluster 3 in blue colour.



6. Discussion

Most of Beijing's shopping centers are located in the “Three Rings” (City Center). The number of shopping malls in Cluster 1 is the largest, but there is only one data, indicating that this is the largest area in Beijing, and the number of shopping malls in Clusters 0 and 3. On the other hand, the number of malls in cluster 2 is very small, and there is no shopping center at all. This is a great opportunity to open a new shopping mall and a potential area, as existing shopping centers have few competitors. Therefore, we recommend Mr. Lee, first of all, do not choose to open a shopping mall in the cluster 1 area, because the competition is very intense. Second, choose to open a shopping center in clusters 0 and 3, but it should also be determined based on various factors such as local land auction and consumption levels. Third, you can choose the cluster 2 area to open a shopping center, you can compete with the original shopping malls in the area, but you need your own shopping malls to be more distinctive, so that you can stand out from the competition.

7. Conclusion

In this project, we have completed the process of identifying business problems and specified required data, extract and prepare data, perform machine learning by clustering data divided into 4 clusters based on their similarity, and finally provide advice to stakeholders, the best place for our good friend Lee to open up new shopping malls. To answer the business questions raised in the introduction, the project's suggestion is that the neighborhood in cluster 2 is the most popular and most promising place to open a mall, because there are no malls. The findings of the project will help Lee take advantage of high-potential locations while avoiding competition with peers. Of course, the specific location of the mall must also consider a variety of factors, we simply provide some candidate locations for Lee from the perspective of location information, and finally hope that our good friend Lee can go well.