IBM Applied Data Science

Capstone Project

Opening a Boba Tea Place in San Fernando Valley

By Scout Zhou

May 29, 2020

Introduction

Background

Boba milk tea or bubble milk tea has been the most trending food in Eastern Asian since early 2000, and it is increasingly getting more attention globally in recent years. Boba culture started as early as in the late 80s in Taipei. While Milk tea has always been well-known in Taiwan and other parts of the world, adding some chewy tapioca balls to traditional milk tea happens to be a revolutionary outbreak in this industry. Nowadays, Boba tea has exploded in popularity across the US. Some famous Boba shop includes CoCo Fresh Tea&Juice, Gong Cha, Yi Fang Taiwan Fruit Tea, and etc.

Business Problem and Target Audience

As mentioned above, the Eastern Asian community, especially the younger Eastern Asian community is profoundly passionate about Boba milk tea. It will be beneficial for property developers to consider taking advantage of this trend and open a Boba milk tea shop. Opening a new Boba milk tea shop will require more consideration than it seems. The profit of having a small business as a Boba milk tea shop will heavily depends on its location. Hence, this capstone project will focus on finding the best geographical location of having a Boba milk tea shop in San Fernando Vally in California. As designed, this project will be particularly helpful for the people who would like to have their own small business and invest in the Boba milk tea industry.

Data and Methodology

The data we needed for this project can be generally sorted into three categories, including

- The names of neighborhoods of the San Fernando Valley will be used to determine the scope of this project. The list of neighborhoods in San Fernando

Valley is provided by Wikipedia. The Beautifulsoup package with python is utilized to scrap the neighborhoods' name from the given website.

- The location of neighborhoods within the scope, including the Longitude and Latitude, will be used for visualization purposes.
- The environment of each neighborhood, especially the venue's information in each region obtained by Foursquare.

The ideal location of a new Boba milk tea shop has to satisfy two essential conditions. Firstly, the Boba milk tea shop will be opened in a region with a higher Asian population since Boba milk tea originated from Eastern Asia, and it is mostly approved by the Eastern Asian group. Following by that, the ideal location is centered or next to crowded regions for better business opportunities. Data science methodology and machine learning techniques such as k-mean clusterings are utilized to analyze the features of each region we investigated for finding the best location of opening a new Boba milk tea shop.

Data Analysis

Data Acquisition and Cleaning

As discussed above, the original neighborhood list is scrapped from the Wikipedia page for San Fernando Valley (https://en.wikipedia.org/wiki/Category:Communities_in_the_San_Fernando_Valley). Furthermore, the name and geographical information of each neighborhood are obtained by using Python packages, Beautifulsoup and Nominatim respectively. After extracting the essential information about the neighborhoods, and assembling all the data into a panda data frame, the first five neighborhoods we will be investigated are displayed in Fig.1, and there are 31 neighborhoods in total. A visual demonstration of all the neighborhoods is provided in Fig 2.

	Neighborhoods	Latitude	Longitude
0	Arleta, Los Angeles	34.241327	-118.432205
1	Burbank, California	34.181648	-118.325855
2	Calabasas, California	34.144664	-118.644097
3	Canoga Park, Los Angeles	34.201078	-118.597826
4	Chatsworth, Los Angeles	34.259571	-118.602325

Fig 1. First Five Neighborhoods displayed with a corresponding geographical location

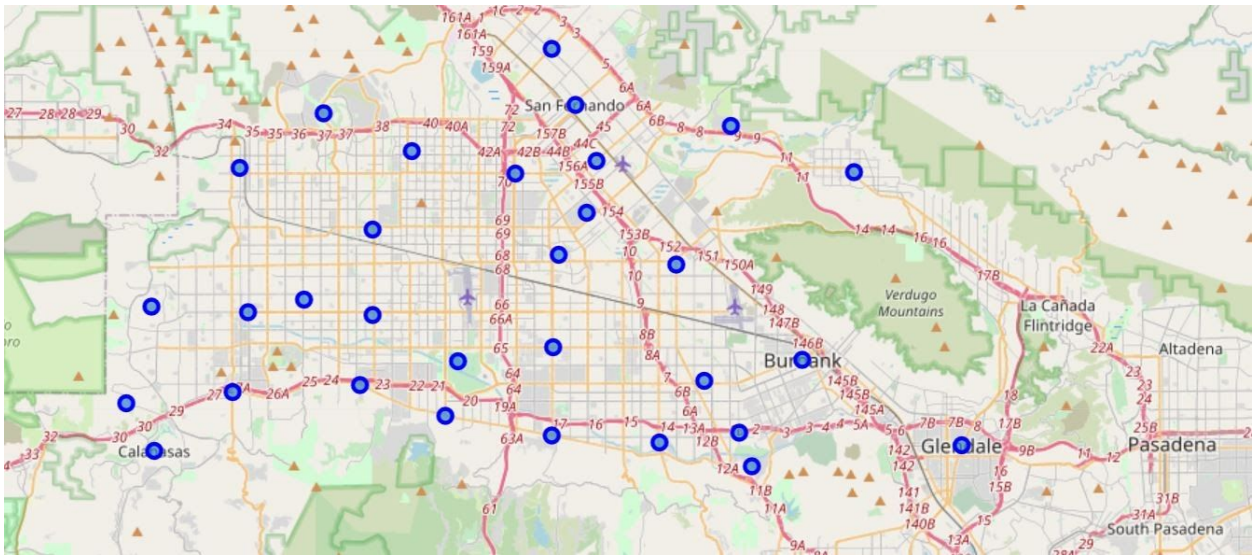

Fig 2. A visual demonstration of all the neighborhoods

Exploratory Data Analysis

There are two features we are seeking for the best location of a new Boba milk tea shop, including a higher Asian population and locating close to crowded areas. In order to further exploring the surrounding of the neighborhoods. The main and only approach of this part we will use is acquiring the data about all the venues in the neighborhoods. The venue's data are provided by Foursquare API. For each neighborhood, we acquire the details about the top 50 venues located within 1000

meters of the center of each neighborhood by passing the geographical coordinates of the neighborhoods to Foursquare API. Furthermore, the most popular venues for each investigated region. For this project, the size of Asian population in each region is evaluated by the number of Asian Restaurants. The highest-ranked neighborhoods are listed in the following figure.

	Neighborhoods	Asian Restaurant
182	North Hollywood, Los Angeles	1
234	Northridge, Los Angeles	1
577	Woodland Hills, Los Angeles	1
20	Canoga Park, Los Angeles	1
74	Glendale, California	1
220	Northridge, Los Angeles	1
245	Panorama City, Los Angeles	1
478	Toluca Lake, Los Angeles	1
15	Canoga Park, Los Angeles	1
413	Sunland-Tujunga, Los Angeles	0
409	Sunland-Tujunga, Los Angeles	0
410	Sunland-Tujunga, Los Angeles	0
411	Sunland-Tujunga, Los Angeles	0
412	Sunland-Tujunga, Los Angeles	0
0	Arleta, Los Angeles	0

Fig 3. The list of the regions with higher Asian population

Finally, the neighborhoods are clustered by k-mean clustering based on the venue's information in order to determine whether a neighborhood is a business active region. The neighborhoods are divided into 7 clusters depends on the categories of their most popular venues. The following map demonstrates the distribution of each cluster in San Fernando Valley. The crowded region are marked with red or orange dot while the rest neighborhoods are not.

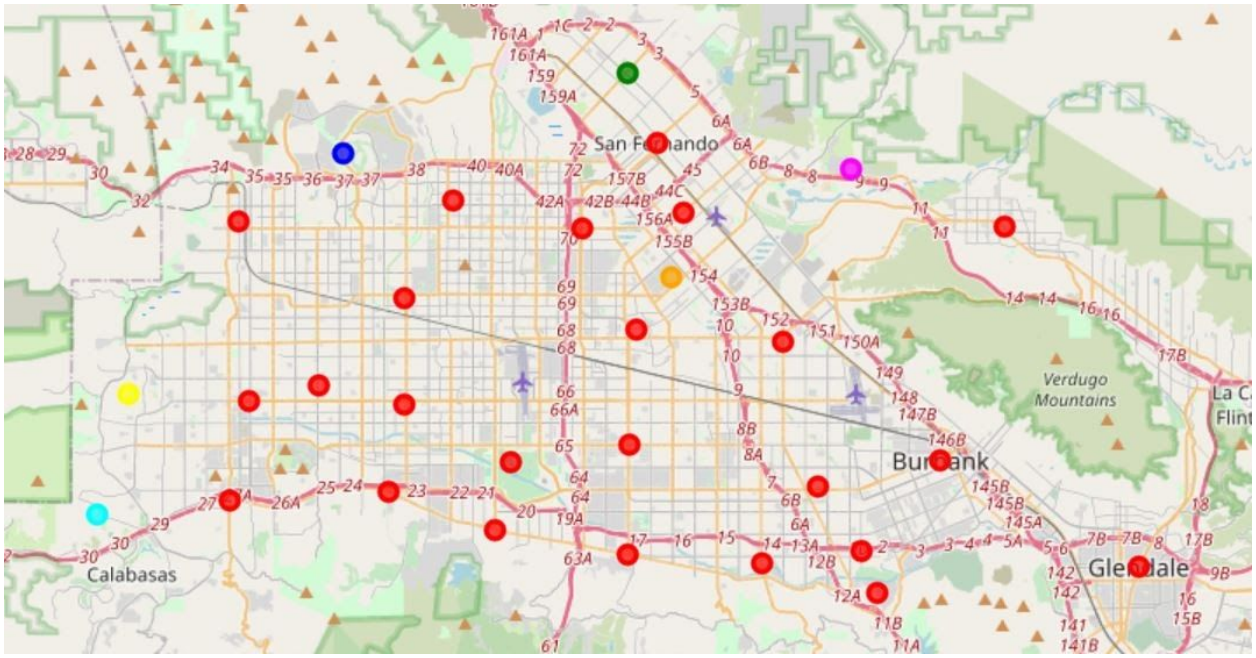

Fig 4. A map of neighborhoods clustered into 7 groups based on their most popular venues. The neighborhoods marked with red and orange dots are the crowded regions while the rest are rather isolated.

Discussion

For this project, the neighborhoods in San Fernando Valley are examed on the basis of two conditions. However, according to the data analysis, most parts of San Fernando Valley are busy and livid regions. Hence the demand for a higher Asian population is more significant for seeking the best location of opening a Boba milk tea shop. In addition to that, since the San Fernando Valley is not a very diverse region, we could improve our analysis process by having less clusters.

Conclusion

Based on the investigation of all the neighborhoods in the San Fernando Valley, the data indicates that most parts of the San Fernando Valley are either centered or located next to the livid regions. Hence opening a Boba milk tea shop in any region with

a higher Asian population will be an excellent choice. Those neighborhoods include North Hollywood, Northridge, Woodland Hills, and etc.

Reference

A brief history of Boba Milk tea
https://www.foodandwine.com/tea/bubble-tea-taiwanese-street-drink-turned-american-addiction

Lab: Segmenting and CLustering Neighborhoods in New York City
https://labs.cognitiveclass.ai/tools/jupyterlab/lab/tree/labs/DP0701EN/DP0701EN-3-3-2-Neighborhoods-New-York-py-v1.0.ipynb

Wikipedia page-San Fernando Valley
https://en.wikipedia.org/wiki/Category:Communities_in_the_San_Fernando_Valley