Setting Up a New Coffee Shop in Los Angeles, US

Tien Nguyen

1. Introduction

Coffee is one of those drinks that a lot of people can't do without. A good portion of the public starts their day with coffee. Many of them continue to drink it well past lunchtime and all through the day. According to the National Coffee Association USA, more than 450 million cups of coffee are consumed in the United States every day and as much as 63% of American adults drink coffee daily. People have their coffee in two ways: making their own coffee at home or going to a coffee shop. There are a lot of reasons people love to go to the coffee shop. Apart from getting a cup of coffee with the high-quality ingredients, best brewing recipes, consistency, fresh and appealing sweet & savory selections, people want to go to coffee houses because they want to meet up or gossip with your friend, do some work, read a book, entertain or simply to pass some time. Seeing that, many businesses want to open a new coffee shop or expand their coffeehouse chain in the future.

Opening a coffee shop can be extremely profitable if we do it right. There are many factors that make a successful coffee shop. Location is one of them. In this project, we want to focus on the Los Angeles area only because it is the largest city in California, United States which means it is also the most populated city in California. With that amount of population, Los Angeles is considered as a good place to start this business, but the question is "which neighborhoods in Los Angeles should a new coffee shop be located ?"

In this project, data science methodology and one of the machine learning techniques(clustering) are used to assist our contractors/clients in finding a neighborhood that their coffee shop should be set up in.

2. Data Sets

There are two datasets in the project: neighborhoods in Los Angeles dataset and Foursquare venue dataset. The neighborhoods in Los Angeles dataset is scrapped directly from the Wikipedia webiste while the Foursquare venue dataset is crawled directly from Foursquare API.

To get the neighborhoods in Los Angeles, I used the BeautifulSoup package to be able to scrape the Wikipedia website and then got the geographical coordinates of the neighborhoods using the neighborhood Wikipedia pages along with the Python Geocoder package which will give me the latitude and longitude coordinates of the neighborhoods. As a result, the Los

Angeles dataset has 190 neighborhoods. For each neighborhood, the data consists of the neighborhood's latitude and longitude.

	Neighborhood	Latitude	Longitude
0	Angelino Heights, Los Angeles	34.070278	-118.254722
1	Angeles Mesa, Los Angeles	33.994200	-118.313600
2	Angelus Vista, Los Angeles	34.046954	-118.317488
3	Arleta, Los Angeles	34.241944	-118.425556
4	Arlington Heights, Los Angeles	34.042222	-118.318889

Fig 1: An excerpt from the neighborhoods in Los Angeles dataset

After having the neighborhood's latitude and longitude, I passed them to Foursquare API to get the top 200 venues that are within a radius of 2000 meters for those neighborhoods. As a result, the Foursquare venue dataset has 16909 observations. Each observation consists of neighborhood latitude, neighborhood longitude, venue, venue category.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Category
0	Angelino Heights, Los Angeles	34.070278	-118.254722	Guisados	Taco Place
1	Angelino Heights, Los Angeles	34.070278	-118.254722	Halliwell Manor	Performing Arts Venue
2	Angelino Heights, Los Angeles	34.070278	-118.254722	Eightfold Coffee	Coffee Shop
3	Angelino Heights, Los Angeles	34.070278	-118.254722	Subliminal Projects	Art Gallery
4	Angelino Heights, Los Angeles	34.070278	-118.254722	Button Mash	Arcade

Fig 2: An excerpt from the neighborhoods in Los Angeles dataset

In addition, the population in a neighborhood also plays an important role when deciding if we should open a new coffee shop there. However, there is no such good website for me to get the population in each neighborhood. Therefore, my assumption is that the more a neighborhood has venues, the more residents live in that neighborhood. Then I classified the population into three groups, depending on the number of venues that each neighborhood has.

- Small population, with the number of venues between 0 and 60
- Medium population, with the number of venues between 61 and 100
- Large population, with the number of venues of 100 or more

	Neighborhood	Total Venue	Population Classification
0	Angeles Mesa, Los Angeles	75	Medium
1	Angelino Heights, Los Angeles	100	Medium
2	Angelus Vista, Los Angeles	100	Medium
3	Arleta, Los Angeles	52	Small
4	Arlington Heights, Los Angeles	100	Medium

Fig 3: An excerpt from the additional venue and population classification table

From the 3 tables above, the geographical coordinates are used to plot the map of Los Angeles with the neighborhoods while venue data is used to perform clustering on the neighborhoods. Because the venue category and population classification column are both categorical, I used one-hot encoding to convert them to numeric. Besides, our aim is to set up the new coffee shop, so I deleted the other venue categories (only kept the coffee shop column.) The Coffee Shop column represents the number of coffee shops in each neighborhood.

	Neighborhood	Small	Medium	Large	Coffee Shop
0	Angeles Mesa, Los Angeles	0	1	0	1
1	Angelino Heights, Los Angeles	0	1	0	13
2	Angelus Vista, Los Angeles	0	1	0	7
3	Arleta, Los Angeles	1	0	0	0
4	Arlington Heights, Los Angeles	0	1	0	6

Fig 4: The final table for performing clustering on the neighborhoods