IBM APPLIED DATA SCIENCE CAPSTONE

Opening a New Chinese Restaurants in Los Angeles, California.

by:
Stephen Kofi Acheampong

June 12, 2020

Introduction

Los Angeles, with one of the largest diasporic Chinese populations outside Asia, is widely regarded as an epicenter for Chinese cuisine. Although the best chinese restaurants are mostly concentrated in the ethnic enclaves of the San Gabriel Valley, there are increasingly more in neighborhoods throughout L.A., from historic Chinatown to the Westside.
Combining exotic flavors with great services and availability, Chinese food slowly climbed the rungs of the ladder to emerge as the top ethnic cuisine, and the most popular worldwide.
For chinese restaurants owners are taking advantage of this trend to build more restaurants to cater for the rising demand.
As a result, there are many chinese restaurants in the city of Los Angeles. But as with any business decision, opening a new chinese restaurants requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the restaurants is one of the most important decisions that will determine whether the chinese restaurants will be a success or a failure.

Business Problem

The objective of this capstone project is to analyse and select the best locations in the city of Los Angeles, California to open a new chinese restaurants. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In the city of Los Angeles, California, if a chinese restaurants owner or franchisee is looking to open a new chinese restaurants, where would you recommend that they open it?

Target Audience of this project

This project is particularly useful to chinese restaurants owners or franchisee looking to open or invest in new chinese restaurants in Los Angeles, California. This project is timely as a lot of people in the city are craving for chinese restaurants and more of such can benefit both investors and customers immensely.

Description of Data

Data of different venues in different cities in Los Angeles from Wikipedia will be used. In order to gain that information we will use "Foursquare" locational information to get a list of all venue and filtered to get a list of chinese restarants in such locations.
We will need the following data:

- List of neighbourhoods in Los Angeles. This defines the scope of this project which is confined to the city of Los Angeles, California.
- Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and also to get the venue data.

- Venue data, particularly data related to chinese restaurants. We will use this data to perform clustering on the neighbourhoods.

Sources of data and methods to extract them.

This Wikipedia page (click to visit link) contains a list of 88 neighbourhoods in Los Angeles. We will use web scraping techniques to extract the data from the Wikipedia page. Then we will get the geographical coordinates of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods.
After that, we will use Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the chinese restaurants 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), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium).

Methodology

Firstly, we need to get the list of neighbourhoods in the city of Los Angeles. Fortunately, the list is available in the Wikipedia page (click to visit link). We will do web scraping using Python pandas to extract the list of neighbourhoods data into a pandas dataframe. 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 Foursquare API. 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 dataframe and then visualize the neighbourhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Los Angeles. 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 neighbourhoods 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 "chinese restaurants" data, we will filter the "chinese restaurants" 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

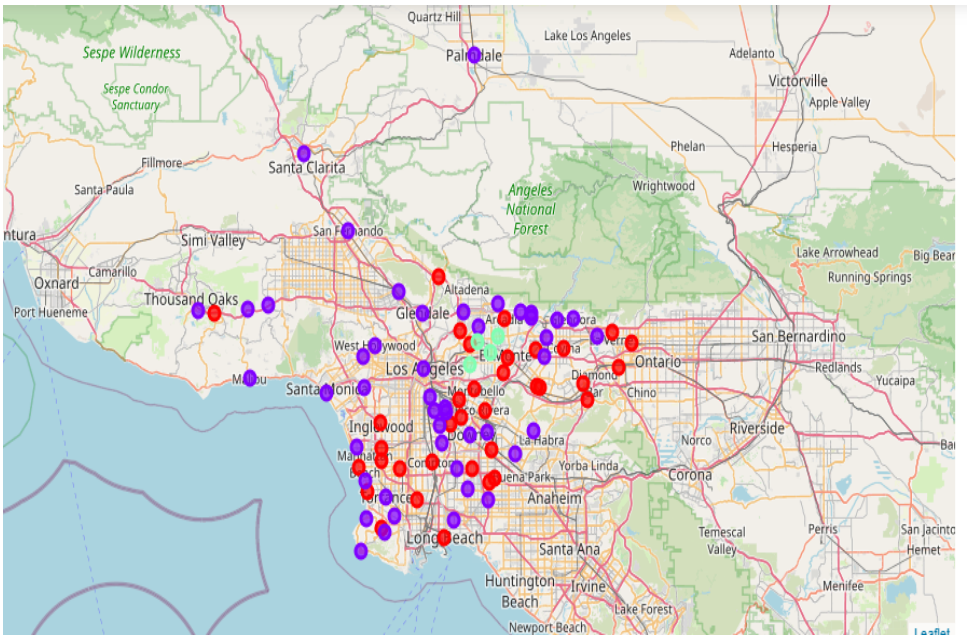
cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 3 clusters based on their frequency of occurrence for "chinese restaurants".
The results will allow us to identify which neighbourhoods have higher concentration of chinese restaurants while which neighbourhoods have fewer number of chinese restaurants. Based on the occurrence of chinese restaurants in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new ones.

Results

The results from the k-means clustering show that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for "chinese restaurants":

- Cluster 0: Neighbourhoods with moderate number of chinese restaurants.
- Cluster 1: Neighbourhoods with the highest number existence of chinese restaurants.
- Cluster 2: Neighbourhoods with the least number existence of chinese restaurants.

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

As observations noted from the map in the Results section, most of the chinese restaurants are concentrated in the central area of Los Angeles city, with the highest number in cluster 1 and moderate number in cluster 0. On the other hand, cluster 2 has very low number to no chinese restaurant in the neighbourhoods. This represents a great opportunity and high potential areas to open new chinese restaurants as there is very little to no competition from existing chinese restaurants.
Meanwhile, chinese restaurants in cluster 1 are likely suffering from intense competition due to oversupply and high concentration of chinese restaurants. From another perspective, the results also show that the oversupply of chinese restaurants mostly happened in the central area of the city, with the suburb area still have very few chinese restaurants. Therefore, this project recommends chinese restaurants owners or franchisee to capitalize on these findings to open new chinese restaurants in neighbourhoods in cluster 2 with little to no competition.
Lastly, they are advised to avoid neighbourhoods in cluster 1 which already have high concentration of chinese restaurants and suffering from intense competition.

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 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders that is chinese restaurants owners or franchisee regarding the best locations to open a new chinese restaurant.
To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 2 are the most preferred locations to open a new chinese restaurant. 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 chinese restaurant.