IBM Capstone Project

1. Introduction

1.1 Background:

Los Angeles has one of the greatest Asian populations in the US. One of the favourite go-to desserts/drinks of the younger Asian population is boba, also known as bubble tea. From my experience living in Los Angeles, it is not always easy to get boba if you do not live in "Asian neighborhoods". With poor public transportation and a driving culture, getting bubble tea could be a 30-minute or longer venture. Therefore, there are opportunities to open more boba shops in Los Angeles where they are not already saturated.

1.2 Business Problem:

As there are opportunities to open more boba shops to capture the demand of those who do not live close to areas already filled with bubble tea options, this project will focus on identifying neighborhoods where there are 1) a few, 2) a moderate amount, and 3) a lot of boba shops. This will give a preliminary view on where new bubble tea shops could thrive!

1.3 Target Audience:

A lot of boba shops are franchise businesses, meaning independent entrepreneurs can use the rights to the bubble tea shop's business name, logo, and products to operate an individual location. The target audience of this analysis will be these individuals who want to open a new location for these bubble tea franchises but are wondering what a good location would be.

2. Data

We will be using the following data:

- List of neighborhoods in Los Angeles: To define the scope of the project and provide the basis for clustering
  - From: https://en.wikipedia.org/wiki/List_of_districts_and_neighborhoods_of_Los_Angeles (https://en.wikipedia.org/wiki/List_of_districts_and_neighborhoods_of_Los_Angeles)
- Latitudes and longtitudes of those neighborhoods: To plot the map and to get venue data
  - From: Geocoder
- Venue data: Particularly those related to bubble tea shops, to perform clustering on neighborhoods
  - From: Foursquare API
  - This data include information such as venue names, their coordinates and their categorie

3. Methodology

3.1 Scrape Data

First, scrape data from the Wikipedia link by using Beautiful Soup and append the data into a list. Due to the fact that the data is from Wikipedia, the neighborhood names can include footnotes label such as [1] and [2] next to them. Those will be cleaned using re. The cleaned list will then be transformed into a pandas dataframe

3.2 Find Coordinates of Neighborhoods and Create a Map

Geocoder is used to locate the coordinates of the neighborhoods in the dataframe. The coordinates are then added into the dataframe as new columns, creating this dataframe with 3 columns which are 'neighborhoods', 'Latitude' and 'Longitude'.

Next, using folium, a map of LA is created. Markers are then added to the map by referencing the the dataframe.

3.3 Explore and Analyse Neighborhood

In order to explore the neighborhood, we made calls using the Foursquare APIs. For each neighborhood, we are able to get venues information including their names, coordinates, and categories. This information is captured into a dataframe.

Next, we took a look at the unique venues categories to explore what kind of venues exist in the neighborhood. In order to get this view into a table, we shape the table with neighborhoods still in each row, and venue categories in columns. By grouping the neighborhoods, now we can see how many venues are there in each categories per neighborhood..

Since we are looking at bubble tea shops in particular, we extract the column with the information of bubble tea shops, and show the number of shops per neighborhood.

3.4 Clustering and Create a Map

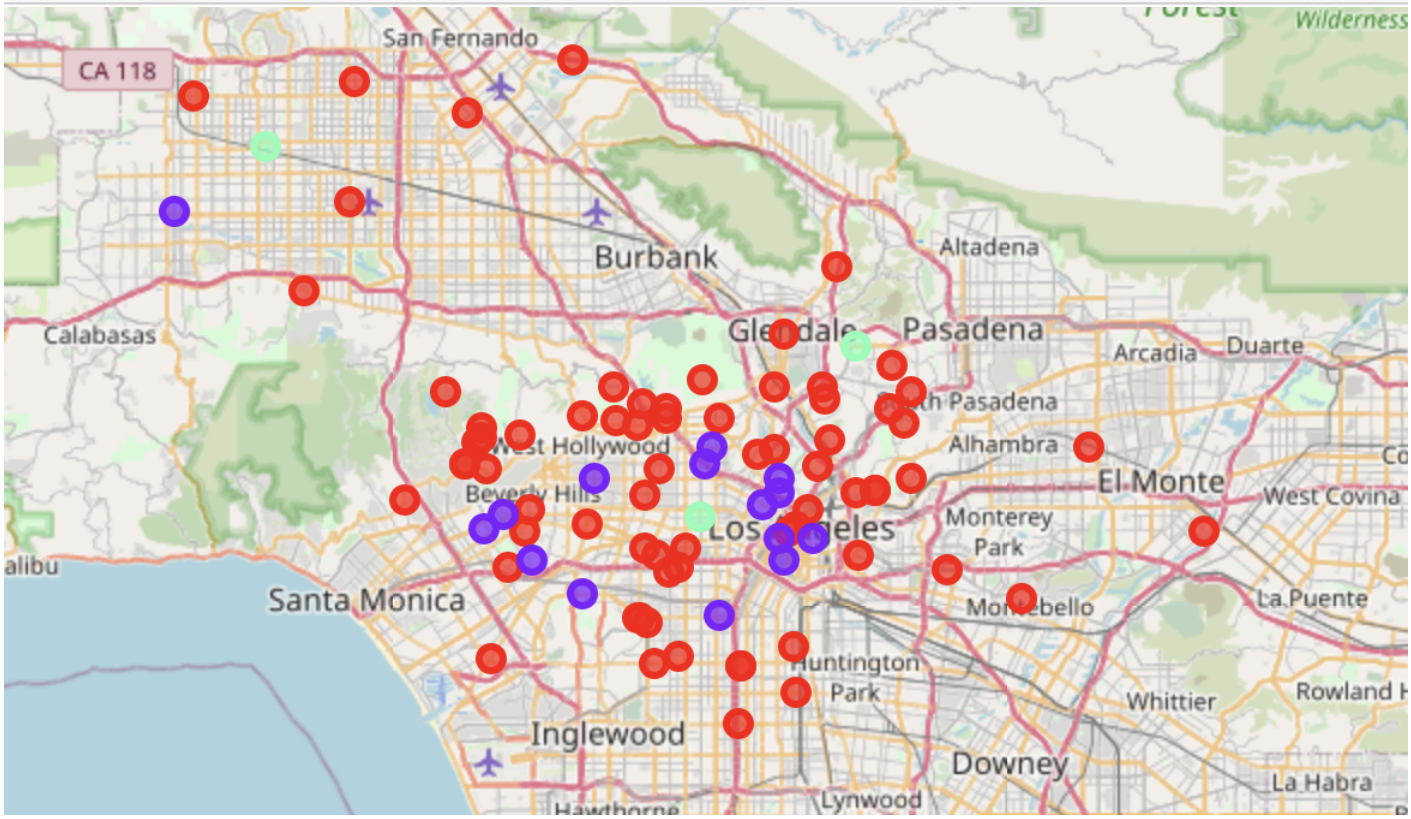
With the bubble tea shop per neighborhood dataframe, we run the k-means clustering specifying 3 clusters. It is meant to show 3 tiers of neighborhoods with high, medium and low amount of bubble tea shops.

We then visualize which of the three clusters these neighborhoods belong to in a map using folium again.

4. Results

K-means clustering divides the neighborhoods into three clusters

- Cluster 0: In red, with low amount of bubble tea shops
- Cluster 1: In purple, with medium amount of bubble tea shops
- Cluster 2: In green, with high amount of bubble tea shops



5. Discussion

According to the results, those who want to open franchises might want to avoid cluster 3 and focus on on cluster 1 and 2 which have fewer boba shops where the market is not saturated yet

However, there is also limitations on this result. Just the number of bubble team shops alone is not the best indicator of whether the market is saturated. We only looked into one attribute which represents the supply but did not look into the demand. It is possible that even though cluster 3 has a lot of bubble tea shops already, that they can still have a high demand for bubble tea that is not satisfied. A next step could be looking at the demand for bubble tea in these areas and compare them with the supply to see what neighborhoods present the best opportunities.

6. Conclusion

In this project, we use neighborhood data from Wikipedia, coordinates from Geocoder, and venue data from Foursquare API to explore where bubble tea shops are located in LA. Then, by using k-means clustering and folium, we are able to find out the groups of neighborhoods with different volumes of bubble team shops and visualize that in a map.

By analyzing data, we are able take a first step in answering a business problem, which is where opportunities lie in opening bubble tea shops.