Capstone Project - The Battle of Neighborhoods

1. Introduction

1.1 Background

San Francisco is one of the financial, cultural and technology centers on the west coast, it has population close to 900,000. San Francisco is a diverse and culturally rich city, where you expect people to live in different lifestyles across different neighborhoods. Indeed, some of its neighborhoods are cozy and relaxing, while some others are busy and commercialized. Therefore, it is very important for either a business to choose where to open a new store or a person to pick where to live with his or her lifestyle.

1.2 Business Problem

I want to utilize Foursquare location to identify venues within each neighborhood, and then use venues' frequencies within each neighborhood to create clusters that provide insightful information for business and people to choose the target neighborhoods to open a new business or live with a desired lifestyle.

2. Data Acquisition and Cleaning

2.1 Data Sources

The data is from the following sources:

a. San Francisco neighborhood list: Wikipedia SF Neighborhoods;
b. Location data: Opencage Geocoder;
c. Venues data: Foursquare;

2.2 Data Cleaning

a. For San Francisco neighborhood data, I used "mw-headline" class in Beautiful Soup Python library to extract the neighborhood list from the Wikipedia website;

	Neighborhood
0	Alamo Square
1	Anza Vista
2	Ashbury Heights
3	Balboa Park
4	Balboa Terrace
...	...
114	West Portal
115	Western Addition
116	Westwood Highlands
117	Westwood Park
118	Yerba Buena

119 rows × 1 columns

b. For the location data, I used the Opencage Geocoder Python library and free API key from Opencage to create a data frame that includes the latitudes and longitudes for each of the neighborhood in San Francisco. I have notice the data frame generated

from the Opencage Python library are not 100% accurate, so I deleted the rows that contained inaccurate location data;

	Neighborhood	lat	lng
0	Alamo Square	37.776360	-122.434688
1	Anza Vista	37.780836	-122.443149
2	Ashbury Heights	37.775599	-122.448068
3	Balboa Park	37.721427	-122.447547
4	Bayview	37.728889	-122.392500
...	...	...	...
101	West Portal	37.741141	-122.465634
102	Western Addition	37.779559	-122.429810
103	Westwood Highlands	37.725726	-122.458199
104	Westwood Park	37.725726	-122.458199
105	Butchertown	37.784827	-122.727802

106 rows × 3 columns

c. Once I had the location data, I used Foursquare to generate 200 venues with a radius of 1,000 meters from the coordinate for each neighborhood.