Capstone Project - The Battle of Neighborhoods:

Week 4 Assignment

Title: Comparing Neighborhoods in LA to SF an NY

1- Problem Description and Background Discussion

A producer of deli meat located in California supplies various types of cured meat to restaurants, supermarkets and sandwich shops located in the Los Angeles metropolitan area. His business is thriving and is considering expanding to another populous city with a similar profile of business distribution and people preferences. Since his business model takes advantage of the appetite for cured meats in Mexican and Italian restaurants and would benefit from expanding to a city with a similar distribution of venues.

To help this producer decide between San Francisco and New York City, we will carry out an analysis of the type of venues and their distribution in each city and determine which of the two shares more similarity with Los Angeles, hence presenting a better chance of success.

We will first collect data on the neighborhoods of all three cities, including name, zip code and geographical coordinates. Then we will extract information on the venues located within each neighborhood. A clustering of the neighborhoods based on the type of venues they contain will follow, together with analysis of the types of venues and plots of the cluster distribution on the map. Finally, we will estimate the similarity between cities to provide a recommendation.

2- Data Description and Application for Solving the Problem

The extraction and processing of the source data followed several steps:

a) <u>Web Data Scraping:</u>

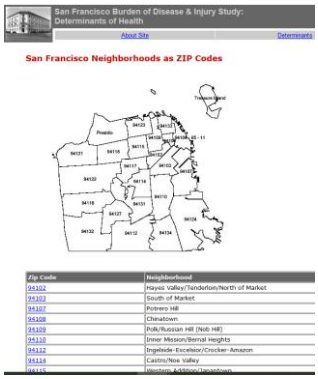
To obtain geographical information from each city, one web site was selected in each case that included a table with at least zip-codes and neighborhood names for that city.

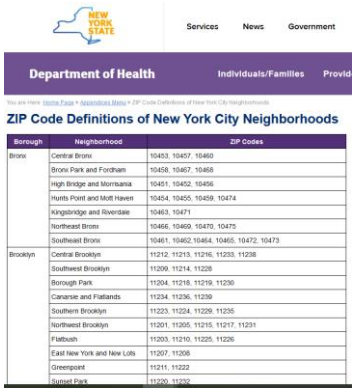
The data was scraped using the library **requests** to grab html data and the library **BeautifulSoup** to scrape html data.

The data for each city was wrangled separately according to the characteristics of the URL source and converted to a dataframe with the columns ‘Neighborhood’ and ‘Zip Code’.

The table below collects the URL information, an image of the site and the resulting dataframe for each of the three cities:

Los Angeles	San Francisco	New York City
http://www.laalmanac.com/communications/cm02_communities.php	http://www.healthysf.org/bdi/outcomes/zipmap.htm	https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoods.htm






```
Number of neighborhoods in LA: 130
```

	Neighborhood	Zip Code
0	Arlington Heights (Los Angeles)	90019
1	Artesia, Artesia (PO Boxes)	90701
2	Athens	90044
3	Atwater Village (Los Angeles)	90039
4	Avalon (PO Boxes)	90704

```
Number of neighborhoods in SF: 21
```

	Zip Code	Neighborhood
0	94102	Hayes Valley/Tenderloin/North of Market
1	94103	South of Market
2	94107	Potrero Hill
3	94108	Chinatown
4	94109	Polk/Russian Hill (Nob Hill)

```
Number of neighborhoods in NY: 42
```

	Neighborhood	Zip Code
0	Central Bronx	10453
1	Bronx Park and Fordham	10458
2	High Bridge and Morrisania	10451
3	Hunts Point and Mott Haven	10454
4	Kingsbridge and Riverdale	10463

b) <u>Coordinates Extraction:</u>

The coordinates of each neighborhood with zip code was extracted using the function **Nominatim** from the library **geopy.geocoders**, which converts an address into latitude and longitude values. For simplicity and to reduce ambiguity given that some zip codes included more than one neighborhood name, the zip code value was used to obtain the latitude and longitude values. Whenever a zip code did not return valid latitude or longitude coordinates, the entire row was deleted. The data was then added to the dataframe of each corresponding city under the columns 'Latitude' and 'Longitude'.

The table below displays the first few rows of each dataframe with all the necessary geographical information to extract venue information in the next step.

Los Angeles

```
Number of neighborhoods in LA: 125
```

	Neighborhood	Zip Code	Latitude	Longitude
0	Arlington Heights (Los Angeles)	90019	34.047371	-118.336046
1	Artesia, Artesia (PO Boxes)	90701	33.868528	-118.077698
2	Athens	90044	33.981914	-118.287489
3	Atwater Village (Los Angeles)	90039	34.118121	-118.264129
4	Avalon (PO Boxes)	90704	33.341730	-118.328136

San Francisco

```
Number of neighborhoods in SF: 21
```

	Zip Code	Neighborhood	Latitude	Longitude
0	94102	Hayes Valley/Tenderloin/North of Market	37.779491	-122.418224
1	94103	South of Market	37.774425	-122.411091
2	94107	Potrero Hill	37.793634	-122.408295
3	94108	Chinatown	37.791043	-122.406578
4	94109	Polk/Russian Hill (Nob Hill)	37.793815	-122.420597

New York City

```
Number of neighborhoods in NY: 42
```

	Neighborhood	Zip Code	Latitude	Longitude
0	Central Bronx	10453	40.852348	-73.911965
1	Bronx Park and Fordham	10458	40.861569	-73.888765
2	High Bridge and Morrisania	10451	40.828381	-73.927084
3	Hunts Point and Mott Haven	10454	40.807728	-73.918198
4	Kingsbridge and Riverdale	10463	40.884718	-73.887248

c) Venue Extraction:

With the coordinates of each neighborhood of each city collected into each corresponding dataframe, the **FourSquare API** set up in week 1 of the course was employed to extract common venues within a pre-defined radius per neighborhood. Because the neighborhoods in Los Angeles can extend a larger area than those in San Francisco or New York, the search radius was adjusted slightly for each city. In particular, a radius of 1000 meters was selected for Los Angeles while 500 meters was selected for both San Francisco and New York City, respectively.

Using code from previous weeks, only relevant information about each returned venue was collected into three separate dataframes, one per each city.

The table below displays the first few rows of the venue dataframes extracted from FourSquare:

Los Angeles

	Venue	Venue Latitude	Venue Longitude	Venue Category
0	PizzaRev	34.048585	-118.336439	Pizza Place
1	Smart & Final Extra!	34.047692	-118.335932	Grocery Store
2	Planet Fitness	34.047774	-118.338605	Gym / Fitness Center
3	Jersey Mike's Subs	34.048449	-118.337419	Sandwich Place
4	PetSmart	34.048184	-118.335489	Pet Store
5	La Fayette Square	34.043205	-118.333813	Neighborhood
6	Midtown Crossing	34.048047	-118.337077	Shopping Mall
7	Mateo's Ice Cream & Fruit Bars	34.047588	-118.327972	Ice Cream Shop
8	El Compita	34.048592	-118.332846	Mexican Restaurant
9	Panda Express Mid-City	34.048654	-118.337556	Chinese Restaurant

San Francisco

	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Asian Art Museum	37.780178	-122.416505	Art Museum
1	Louise M. Davies Symphony Hall	37.777976	-122.420157	Concert Hall
2	Herbst Theater	37.779548	-122.420953	Concert Hall
3	Philz Coffee	37.781433	-122.417073	Coffee Shop
4	War Memorial Opera House	37.778601	-122.420816	Opera House
5	San Francisco Ballet	37.778580	-122.420798	Dance Studio
6	Ananda Fuara	37.777693	-122.416353	Vegetarian / Vegan Restaurant
7	Siam Orchid Traditional Thai Massage	37.777111	-122.417967	Massage Studio
8	August 1 Five	37.780537	-122.420188	Indian Restaurant
9	War Memorial Court	37.779042	-122.420971	Park

New York City

	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Liberato	40.853744	-73.907966	Latin American Restaurant
1	Accra Resturant	40.853871	-73.908421	African Restaurant
2	Wingstop	40.854093	-73.907899	Wings Joint
3	Bravo Supermarkets	40.853936	-73.914144	Grocery Store
4	Papa John's Pizza	40.852429	-73.908976	Pizza Place
5	Dunkin Donuts	40.853817	-73.908724	Donut Shop
6	Chase Bank	40.854087	-73.907631	Bank
7	Food Dynasty	40.853772	-73.909267	Supermarket
8	Subway	40.853887	-73.907285	Sandwich Place
9	Chase Bank	40.850381	-73.916217	Bank

These dataframes are used in subsequent steps to analyze the neighborhood information and extract insights.