Clustering and Comparing the Neighborhoods of San Francisco and Chicago cities

Abdalrhman Abdalla

July 12, 2020

IBM Data Science Professional Certificate

(Applied Capstone Project)

1. Introduction and Problem Statement

The problem to be addressed in this project is for my friend who has a job offer in the United State. Upon his request I will not mention his real name let's, call him Boss. Boss lives in the Middle East and he works on the tech sector and got a job offer from the same company in US. Boss is still confused which city he is going to reside since he can work remotely, he has two major cities in his mind which they are San Francisco, CA and Chicago, IL and he would like to know which neighborhoods are better according to his living style, Boss habits and requirements are below :

- He goes to gym daily.
- He addicted to coffee.
- He likes different types of food but Indian and Italian are his favorite.
- He wants to be near to parks.
- He wants a clean and well-maintained neighborhood.

2. Data acquisition and cleaning

- **Data Source**

A Neighborhood list of the cities of San Francisco and Chicago is available online in different website and can be found in the following links:

San Francisco

http://www.healthysf.org/bdi/outcomes/zipmap.htm

Chicago

http://www.lizshomes.com/resource_zips.html

Foursquare API's has a database of over 105 million places worldwide and will be consulted for this project. We'll use the Venue Recommendation to explore the two cities, which returns a list of recommended venues near a specific location. To make the query in the Foursquare API we need the coordinates given for a given neighborhood in Latitude and Longitude. So, the first thing I need to do is to get coordinates in each of the cities for each neighborhood.

To get the coordinates given a city name and a Neighborhood we use Geocoder. Geocoder is a geocoding library very easy to use written in python that find locations of addresses, and zip codes.

- **Cleansing**

The data for San Francisco scraped from healthysf website with the list of neighborhoods and the cross-bonding zip code for each neighborhood. In the other hand the data for Chicago neighborhoods were copied from lizshomes website to an excel file and then converted to dataframe after altering a few the neighborhoods names which has multiple zip codes, and . in order to consult nearby venues in each neighborhood the coordinates were added to the two dataframes. So, at the end we end up with a dataframes like this:

	Zip Code	Neighborhood	Latitude	Longitude
1	94102	Hayes Valley/Tenderloin/North of Market	37.777015	-122.421875
2	94103	South of Market	37.772000	-122.408735
3	94107	Potrero Hill	37.760651	-122.394064
4	94108	Chinatown	37.791775	-122.407440
5	94109	Polk/Russian Hill (Nob Hill)	37.790105	-122.420590

San Francisco 1

	Zip Code	Neighborhood	Latitude	Longitude
0	60601	Loop1	41.886255	-87.622310
1	60602	Loop2	41.883250	-87.630795
2	60603	Loop3	41.880890	-87.621270
3	60604	Loop	41.878160	-87.631010
4	60605	Loop, Near South Side	41.869255	-87.626255

Chicago 1