The Battle of Neighborhoods:

A Comparative Analysis of Neighborhoods in San Francisco, CA & Chicago, IL.

Adewale Sanusi

August 2020

1. Introduction/Business Problem.

United States of America is a land of opportunities where many from all over the world want to either visit temporarily or immigrate to permanently and call a new home. As a result of this, many prospective immigrants or newcomers attempt to compare various cities to determine where best suits their needs which could range from job seeking opportunity, business opportunity or a peaceful area to live in etc. We do know that starting life afresh in a new city can be a problem.
This work is a capstone project for the completion of the IBM Data Science Certification, and it will involve a comparative analysis of different neighbourhoods in two different cities. It includes a segmentation of downtown areas, identify various venues and cluster them together. San Francisco, California and Chicago, Illinois are two US cities which has been selected for this project. The analysis will also explore their similarities and differences and the results can be a useful piece of information for new settlers and or prospective immigrants.

2. Data Selection & Description.

Various sets of data will be required to solve the problem highlighted above therefore, the following data sets will be used to complete this project.
Data sets containing;

- United States cities zip codes together with their corresponding geographical coordinates. This will be used to generate data frames for San Francisco and Chicago. (https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude)
- San Francisco zip codes and the area name. (http://www.healthysf.org/bdi/outcomes/zipmap.htm)
- Chicago zip codes and the area names. (https://www.chicagotribune.com/chi-community-areas-htmlstory.html)
- These data containing the zip codes and area names will be merged to create a new data frame for identifying the neighborhoods in both cities which will then be used for segmentation and clustering analysis.

- Foursquare API will be required to generate venues in the selected areas.

3. Methodology

Data scraped or mined from many sources were combined to create data frames containing zip codes and geographical coordinates of different neighbourhoods in both San Francisco and Chicago cities. Python folium library and Foursquare API were also utilized to generate maps for visualization and location venues respectively.

	Zip_Code	City	State	Latitude	Longitude	AreaName
0	94117	San Francisco	CA	37.770937	-122.44276	Haight-Ashbury
1	94131	San Francisco	CA	37.741797	-122.43780	Twin Peaks-Glen Park
2	94114	San Francisco	CA	37.758434	-122.43512	Castro/Noe Valley
3	94107	San Francisco	CA	37.766529	-122.39577	Potrero Hill
4	94116	San Francisco	CA	37.743381	-122.48578	Parkside/Forest Hill
5	94133	San Francisco	CA	37.801878	-122.41018	North Beach/Chinatown

Table. 1: San Francisco data frame

	Zip_Code	City	State	Latitude	Longitude	AreaName
0	60634	Chicago	IL	41.944454	-87.79654	Belmont Cragin, Dunning, Montclare, Portage Park
1	60602	Chicago	IL	41.882937	-87.62874	Loop
2	60601	Chicago	IL	41.886456	-87.62325	Loop
3	60645	Chicago	IL	42.008956	-87.69634	West Ridge
4	60651	Chicago	IL	41.901485	-87.74055	Austin, Humboldt Park

Table. 2: Chicago data frame

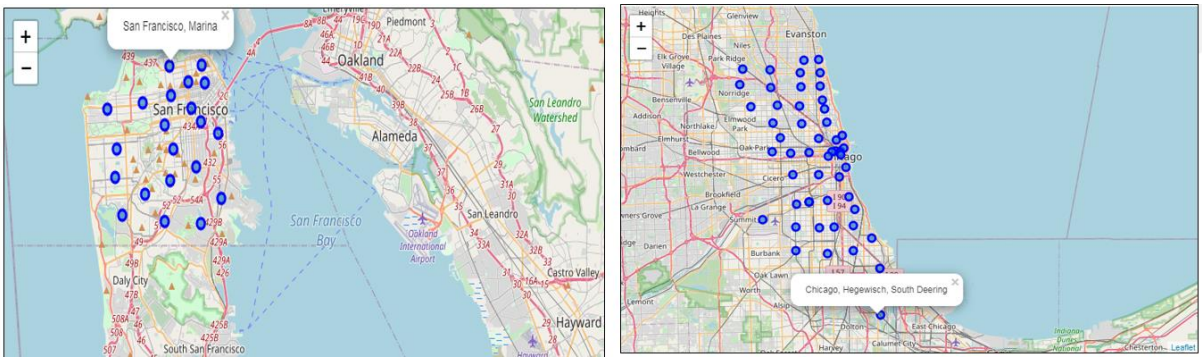

Figure 1. Maps showing area names by zip codes in San Francisco & Chicago cities.

	Zip_Code	City	State	Latitude	Longitude	AreaName	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
0	94103	San Francisco	CA	37.772329	-122.41087	South of Market	1	Nightclub	Gay Bar	Motorcycle Shop	Thai Restaurant	Cocktail Bar	Art Gallery	Restaur
1	94102	San Francisco	CA	37.779329	-122.41915	Hayes Valley/Tenderloin/North of Market	0	Café	Coffee Shop	Hotel	Wine Bar	Theater	French Restaurant	Pizza Place

Table 3. Top 10 venues in Downtown San Francisco.

	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Loop	Hotel	Coffee Shop	Theater	Sandwich Place	American Restaurant	Italian Restaurant	Seafood Restaurant	Pizza Place	Middle Eastern Restaurant	Snack Place
1	Loop, Near South Side	Historic Site	Football Stadium	Park	Athletics & Sports	Sushi Restaurant	History Museum	English Restaurant	Museum	Donut Shop	Parking
2	Loop, Near West Side	Coffee Shop	Sandwich Place	New American Restaurant	BBQ Joint	Mexican Restaurant	Bar	Mediterranean Restaurant	Burger Joint	Italian Restaurant	Donut Shop
3	Loop, Near West Side, Near South Side	Greek Restaurant	Sandwich Place	Coffee Shop	Pizza Place	Café	Gym	Intersection	Spa	Dance Studio	Sports Bar

Table 4. Top 10 venues in Downtown Chicago.

4. Machine Learning: Segmentation & Clustering Analysis

K-means clustering method was used to produce the clusters for downtown San Francisco and Chicago neighbourhoods for ten (10) most common venues.

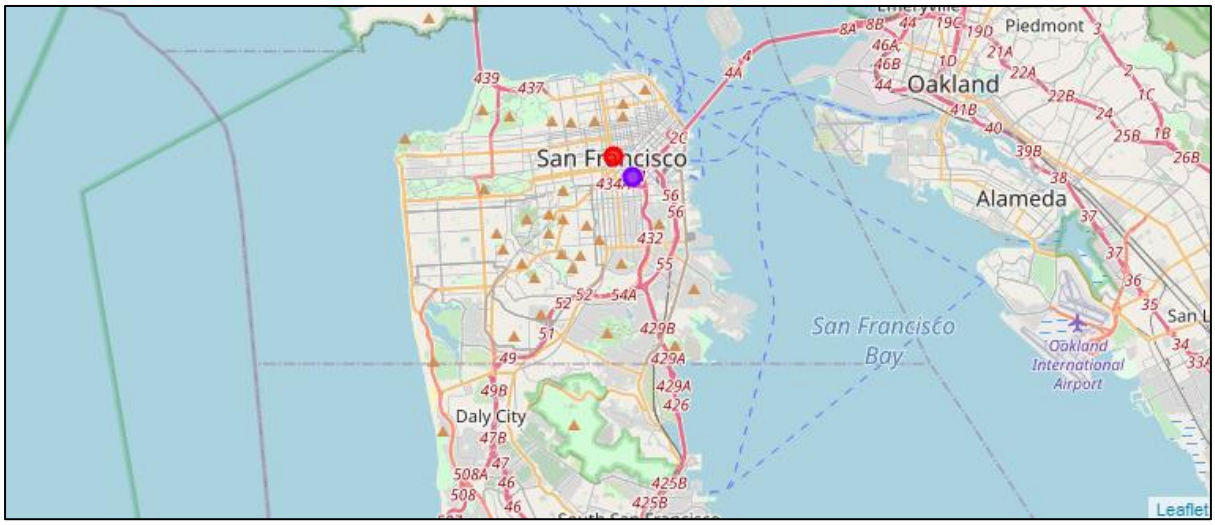

Figure 2. Maps showing downtown San Francisco clusters.

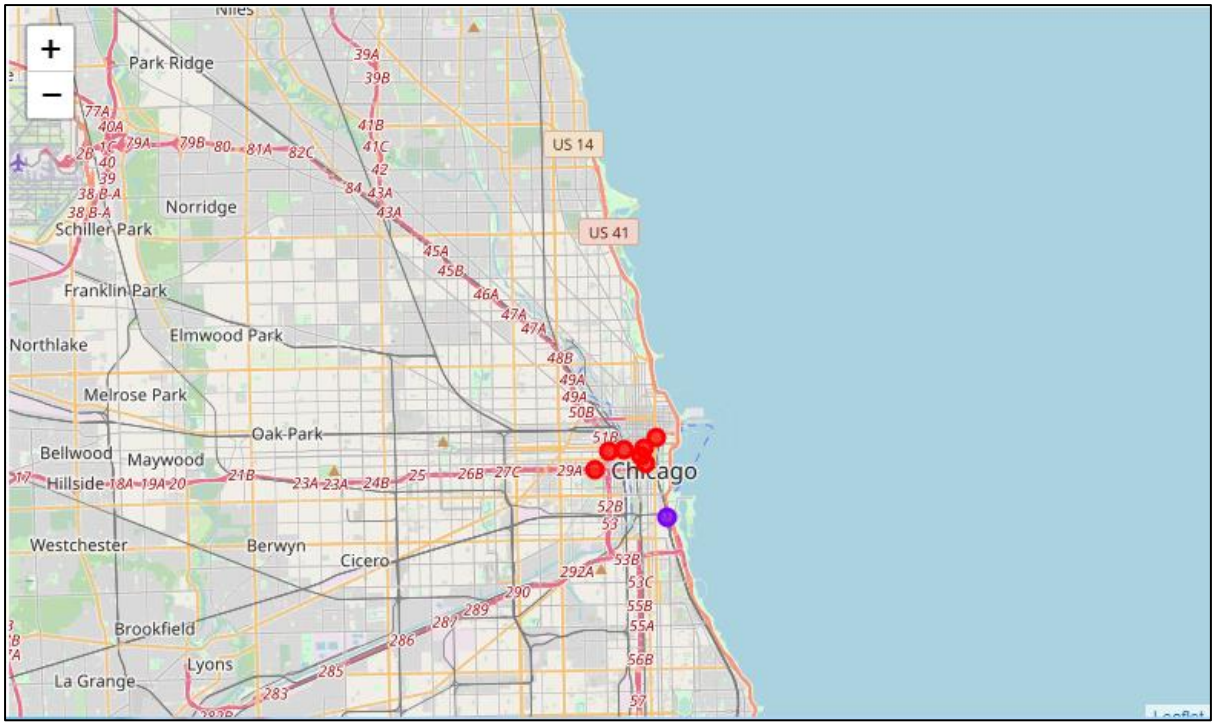

Figure 3. Maps showing downtown Chicago clusters.

5. Discussion.

According to the analyses on both cities data set, it is observed that San Francisco and Chicago downtown areas have a lot of similarities as shown in clusters 1 for both cities. Both clusters (1) are mainly characterized by Hotels, Café or Coffee shops and pizza places which are commonly found in business district area. Therefore, clusters 1 in both San Francisco, California and Chicago, Illinois downtown areas can be referred to as the ***"Business District"***.

Cluster 2 for San Francisco data set, unlike cluster 1, is characterized by relaxation venues such as Nightclub, Gay Bar, Restaurant and lounge. It shows that a lot of leisure activities and relaxations occur in this area. This cluster can be named ***"Home of Relaxation"***. While cluster 2 for Chicago data is characterized by sport centers like football stadium, athletics & sports venue as well as historic or tourist sites such as museum. This cluster can be called the ***"Home of Sport & Tourism".***

6. Summary and Conclusion.

This project made use of data set extracted from different open or public domain sources mainly internet websites. Various python libraries were utilized to fetch, clean, manipulate and visualize the data while foursquare API was used to focus on the venue details of each neighborhood of both San Francisco, California and Chicago, Illinois downtown areas.

Machine learning algorithm was applied for segmentation and clustering analysis to gain more insights on the data.

This analysis has provided us with some good insights and preliminary information on neighborhood

categorization and various activities centers for quick understanding of the cities for newcomers either for job search or to open businesses.

The objectives of the project were met and, with additional data such as crime rates, further comparative analysis can be carried out on both cities which can then be a useful information to newcomers or prospective immigrants who may want to compare both cities either for settling, jobs or other new opportunities.