Comparing Chicago Neighborhoods to San Francisco Neighborhoods

David Drees

December 21, 2020

1. Introduction

1.1 Background

Chicago is the third most populated city in the United States, and is well known for having a diverse community with a variety of restaurant cuisines, social venues, and tourist attractions. It is also comprised of many unique neighborhoods whose residents are used to being around the venues in the neighborhood. Per U.S. Census data, San Francisco is the most common destination for residents that move out of Chicago (about 6.4% of all migrations out of Chicago are to San Francisco). Many people that move from Chicago to San Francisco hope to find a neighborhood to reside in with familiar surroundings.

1.2 Problem

To answer the question of "Which neighborhood should I choose when moving to San Francisco?" I will be exploring the types of venues in each neighborhood of San Francisco and comparing them to the types of venues in each neighborhood in Chicago. This will allow someone to make an informed decision of which new neighborhood to choose when they make the move from Chicago to San Francisco.

2. Data

In order to compare the neighborhoods of Chicago and San Francisco, I needed to gather data from multiple sources. First, I needed to find data on the neighborhoods themselves that would list the neighborhood names as well as their geographical coordinates to be able to explore the types of venues in nearby each one. I was able to download neighborhood data in the form of Geojson files from each city's website.

- Chicago's neighborhood data can be located here

- San Francisco's neighborhood data can be located here

The Geojson files will provide me with boundaries on each neighborhood which will allow me to find what neighborhood a set of coordinates is in, instead of just distance from the center. These files will also give me the ability to create better visualizations such as maps with neighborhood boundaries and choropleth maps.

From there, I would need to call Foursquare's API for each neighborhood to obtain venue data which would be my last source of collected data. After that, I will only need to continue clean and explore the data in order to solve which neighborhoods are most similar.

3. Methodology

3.1 Obtaining Neighborhood Information

After loading the geojson data files into my code, I had to pull out the necessary information in order to make them usable. The files contained a lot of extra information, but I was only focused on the geographic coordinates and the name of the neighborhoods that corresponded with the coordinates. I extracted the necessary data by running a loop through each feature in the file to append neighborhood name and coordinate data to a new DataFrame.

The coordinates that a Geojson file provide are boundary coordinates, and there are often times upwards of 100 sets of coordinates just to create the entire boundary of a neighborhood. Knowing that later, when I call the Foursquare API for each neighborhood to obtain venue data, I will need to enter just one coordinate per neighborhood, I needed to find the center of the boundaries provided. To do this, I took an average of the minimum latitude & longitude and the maximum latitude & longitude in each set of boundary coordinates. The same process was completed for both Chicago and San Francisco data.

3.2 Finding Neighborhood Venue Information

API calls via Foursquare were made in an iteration for every neighborhood in both cities, returning up to the top 100 venues within 500 meters of each neighborhood center. While not all neighborhoods had 100 venues withing 500 meters in Foursquare's database, each had enough information to work with. Considering that I am comparing neighborhoods strictly based on the types of venues that they possess, I created a DataFrame for each city listing all of the categories of venues as columns and each

neighborhood as a feature, with normalized frequency of the category in the corresponding neighborhood labelling each cell.

3.3 "Scoring" Neighborhoods

Once I had normalized venue category data for each neighborhood, I had to make sure the categories matched to be able to compare them. The scope of this model is to take a Chicago neighborhood and find similar ones in San Francisco. For this reason, I dropped all categories of San Francisco venues that did not appear in Chicago venues while on the other hand, I added venue categories in San Francisco that did not originally appear there but did appear in Chicago. The added venues were all given a normalized frequency of 0 for all San Francisco neighborhoods since they did not appear in any venues in the city.

In order to create a score for each neighborhood, I found the linear distance between each category in different neighborhoods. For example, bakeries account for 14% of all venues in Albany Park, Chicago, and they account for 6% of all venues in Chinatown, San Francisco. The linear distance between these neighborhoods is .08 (the absolute value of .14 minus .06). The smaller the number, the more similar the frequency of the category in each city. I then got the sum of the distance of all categories for each neighborhood when compared to a neighborhood in the opposite city. The sum of the distances returned a value between 0 and 2, and again, the smaller the number, the more similar the neighborhood. In an effort to make the results easier to read, I subtracted the score from 2 so that the higher scored neighborhoods would be the most similar. Below is a sample of the score DataFrame.

Neighborhood	Albany Park	Andersonville	Archer Heights	Armour Square	Ashburn
Bayview Hunters Point	0.00000	0.04082	0.00000	0.00000	0.00000
Bernal Heights	0.47500	0.53316	0.25000	0.64537	0.07500
Castro/Upper Market	0.20000	0.62082	0.34000	0.42000	0.12000
Chinatown	0.40000	0.70408	0.24000	0.65630	0.18000
Excelsior	0.55882	0.51441	0.67647	0.47168	0.14706

4. Results

4.1 Visualization

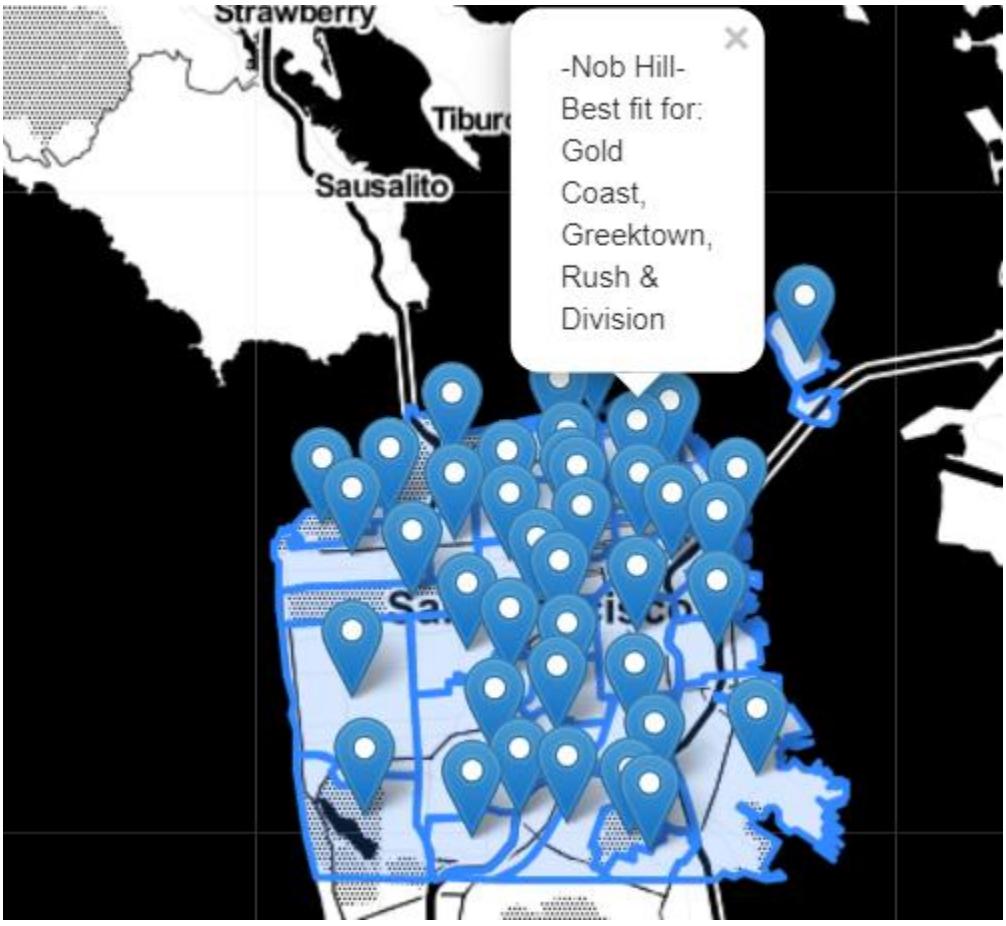
I was able to use all of the information noted above to put together visualizations that make it much easier to view the most similar neighborhoods between cities.

First, I used the scores to create a choropleth map of the neighborhoods in San Francisco that change color based on how similar they are to a neighborhood in Chicago. In the example below, I compared neighborhoods in San Francisco to Gold Coast, Chicago – one of Chicago's most commonly visited tourist destinations and known for its high-end restaurants, luxury accommodations, and proximity to downtown.

The results of the choropleth show the more similar neighborhoods in a darker red color. Gold Coast's most similar neighborhood based on score was Nob Hill. Some familiar with San Francisco's neighborhoods might recognize that Nob Hill is also well known for the same descriptors as Gold Coast, affirming the accuracy of this model.

To provide a more useful, interactive tool for those interested in similarity of other neighborhoods, I provided a map with an overlay of the boundaries of each San Francisco neighborhood as well as a clickable marker that would describe the name of the San

Francisco neighborhood as well as which Chicago neighborhoods it would be the best fit for.



5. Discussion

5.1 Observations

Upon completing this model, I was able to provide "best fit" neighborhoods for anyone in Chicago hoping to find somewhere with similar venues in San Francisco. Recommendations for anyone seeking this information could be found with a simple click in the visualizations in the notebook.

5.2 Future Directions

While I was able to provide neighborhood similarity strictly based on venue categories, there are certainly a myriad of other factors that play a role in similarity of neighborhoods. Neighborhoods could also be compared based on crime rate, population density, proximity to downtown, etc. that would all contribute to making a more informed decision on finding a similar neighborhood. In addition, Foursquares API only allows to

search for venues within a given radius of a coordinate. While this might be an excusable workaround, it might not give the full scope of venue data in a neighborhood. If the comparison were to include all venues within the boundaries of a neighborhood, and possibly even adding weight to the popularity of each venue, it may be possible to achieve more accurate results.

6. Conclusion

In this study, I took venue data from all neighborhoods in both Chicago and San Francisco to answer the question, "Which neighborhood should I choose when moving to San Francisco?" I used the frequency of venue categories in each neighborhood to make a comparison with each other and create a score for each neighborhood. These scores will allow someone to make an informed decision when looking for a neighborhood that brings many of the same characteristics as one they may be used to in Chicago.