A tool for identifying neighborhoods in San Francisco that are similar to neighborhoods in Manhattan

Introduction

Background

Both San Francisco and New York City are centers for tech companies and finance and there are many people that needs to relocate from one city to another due to employment opportunities.

Problem

Because relocation is a complex (and expensive) endeavor, there are companies that helps people in this process. One of the first and most crucial steps in the process is choosing a place of living at the new city. This is not an easy step because many times the customers are not familiar with the neighborhoods at the new city.

In this project we built a tool for a relocation company that is specialized in moving between New York City (specifically Manhattan area) and San Francisco.

The purpose of the tool is to help a customer choosing a neighborhood to live in San Francisco by pointing on the neighborhoods in San Francisco that are the most similar to the customer's favorite neighborhood in Manhattan area.

Interest

Assisting the customer choosing a neighborhood in San Francisco based on his familiarity of neighborhoods in his/her current city will speed up the process of the relocation planning. It will save time and resources by narrowing the area in which the company has to search for real estate, school etc, and make the process of relocation more efficient – meaning more profit to the company and shareholders.

Data

Data sources

In order to start collecting data and ask questions about neighborhoods, we first need to get the locations of the neighborhoods.

San Francisco neighborhoods location

We will use a geojason file containing the boarders of the neighborhoods in San Francisco.

We will have to calculate the center of each neighborhood border in order to get a specific coordinate of the neighborhood.

The geojason file can be downloaded from the following link:

https://cocl.us/sanfran_geojson

Manhattan neighborhoods location

We will use a geojason file containing the neighborhoods locations and names of New-York area.

This geojason file already contains the coordinates of the neighborhoods, however it does not contains information on the boarders of the neighborhoods.

We will use only a subset of this neighborhoods which are localized to Manhattan,

The geojason file can be downloaded from the following link:
https://geo.nyu.edu/catalog/nyu_2451_34572

Foursquare location data

In order to be able to compare between different neighborhoods we will have to retrieve information about them. The Foursquare location data will be used to collect data regarding the venues in the neighborhoods.

Foursquare is a location-based online social network that collects data regarding venues and their location. The data is being collected using more than 30 million people worldwide.

The Foursquare Places API will be used to access to Foursquare's database of venue data.

A link for Foursquare API:

https://api.foursquare.com

Information about the venues in each neighborhood will be retrieved from the Foursquare API. The relative abundance of the different venues types in the neighborhoods will be used to determine the similarities between neighborhoods. The details will be described at the methodology section.

Methodology

In order to compare between neighborhoods in Manhattan and in San Francisco we will have to have the neighborhoods names, and coordinates of the middle of the neighborhoods. This information will be used to retrieve information regarding the neighborhoods via Foursquare.

San Francisco neighborhoods coordinates:

We were using geojason file containing the boarders of the neighborhoods in San Francisco.

In order to get coordinates at the center of the neighborhoods from the geojason file I used the Python package shapely.

Manhattan neighborhoods coordinates:

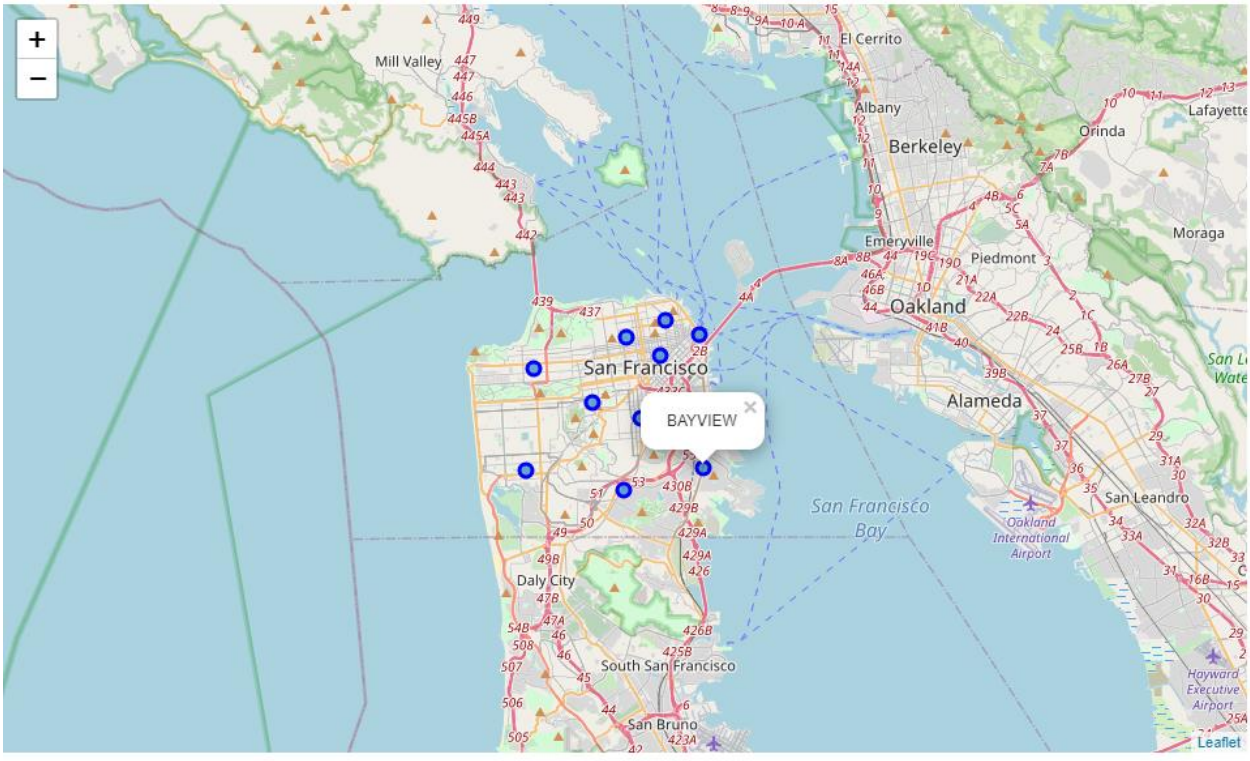
We used a geojason file containing the neighborhood names and coordinates.

The geojason file contained data on neighborhoods in all New-York. We used only the data of neighborhoods in Manhattan.

Map visualization:

For visualization of the neighborhoods on the map, I used the Python package folium.

The neighborhoods centers are represented on the maps as dots, and clicked on – the neighborhood name is presented:



Retrieving venues information:

Foursquare was used to retrieve venues information for each neighborhood in both Manhattan and San Francisco using the neighborhoods coordinates.

For each neighborhood up to 100 venues in radios of 500 meter from the coordinate of the neighborhood was retrieved.

Data cleaning:

The venues types were transformed into dummies and grouped by neighborhood.

Then, the mean of the frequency of occurrence of each venue category was taken.

Finding the most simmilar neighborhoods in San Francisco for each neighborhood in Manhatten using k-nearest neighbors (statistical model)

We iterate over the dataframe of Manhatten neighborhoods, and for each neighborhood we test what are the most similar neighborhoods in San Francisco. For this we **k-nearest neighbors** algorithm.

k-nearest neighbors algorithm group together the most similar neighborhoods based on the venues types abundance.

The clusters numbers for each neighborhood in Manhatten and the neighborhoods in San Francisco where collected in one dataframe.

Finding the most similar neighborhood in San Francisco giving a specific neighborhood in Manhattan

We ask the user to provide a specific neighborhood in Manhatten.

Since we already calculated all of the different combinations of neighborhoods and stored it in a database, we retrieve the relevant data from the dataframe and present on the map only the neighborhoods that cluster together with the neighborhood the user had chosen.

Finding the San Francisco neighborhoods that where more frequently clustered with Neighborhoods from Manhatten

To find which neighborhoods are more similar to neighborhoods in Manhattan, we counted how many times they where clustered together with a neighborhood from Manhattan.

This was done by transforming all the neighborhood with the same cluster number as the neighborhood from Manhattan into 1 (only the neighborhoods that clustered together with it)**,** and all the other neighborhoods into -1. Then the numbers 1 where counted for each neighborhood.

Results

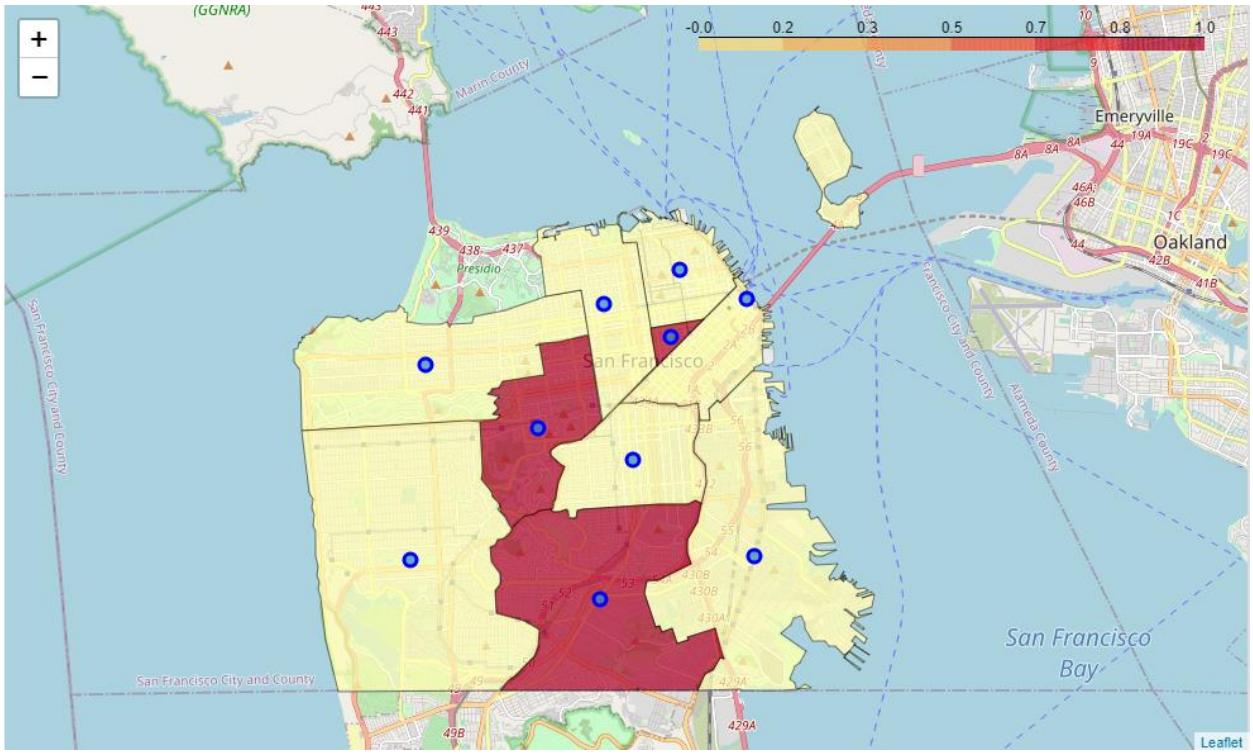
For each neighborhood in Manhattan that the user chooses we get a different set of neighborhoods in San Francisco that are simillar to that neighborhood based on venues types.

Here is an example of a user query for the Manhattan neighborhood: Upper West Side

The neighborhoods in San Francisco that clustered together are:

```
PARK
INGLESIDE
TENDERLOIN
```

In addition to the names, the neighborhoods are marked on San Francisco map in red color according to their boarders:



Exploratory data:

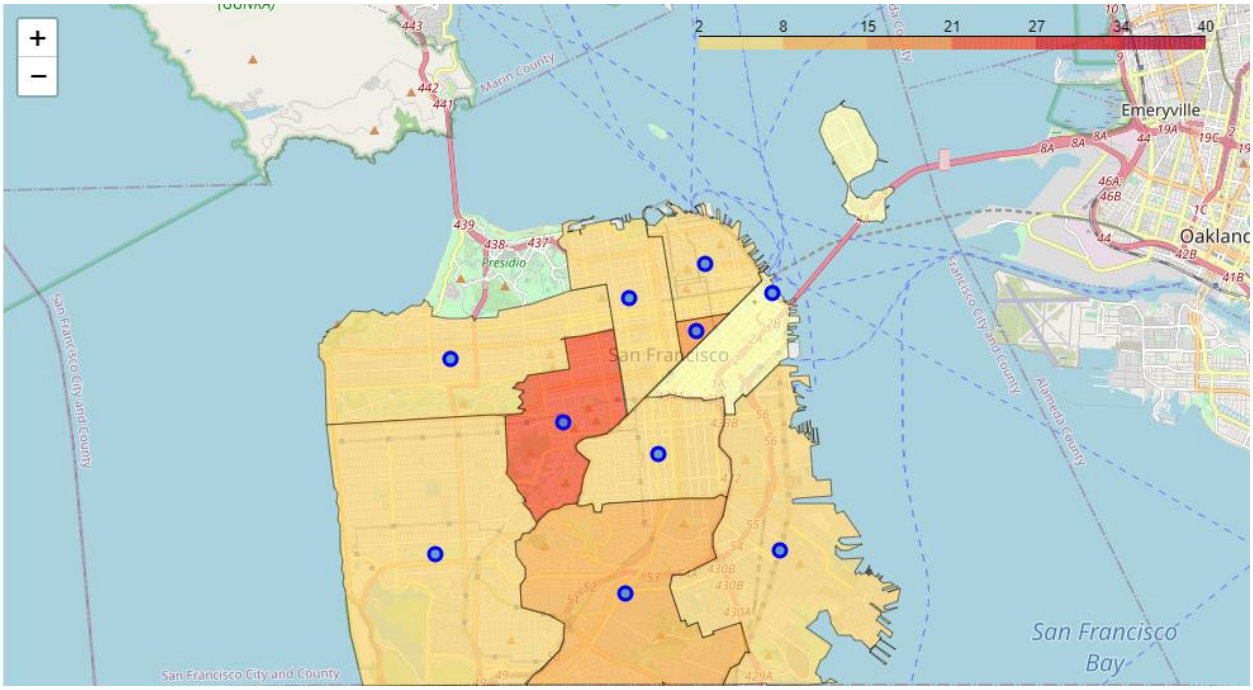
Finding the San Francisco neighborhoods that where more frequently clustered with Neighborhoods from Manhatten

The number of times that each San Francisco neighborhood was clustered together with a neighborhood from Manhattan was calculated.

The neighborhoods were sorted so that the higher occurrence neighborhoods are on the top:

Neighborhood	#
PARK	22
TENDERLOIN	19
INGLESIDE	11
TARAVAL	7
BAYVIEW	5
RICHMOND	4
MISSION	4
CENTRAL	4
NORTHERN	2

The number of times that the neighborhoods were clustered together with Manhattan neighborhoods were also presented on San Francisco map, where higher number represented as more toward red color and lower numbers more toward yellow colors.



Discussion

Given a query neighborhood in Manhattan our tool finds similar neighborhoods in San Francisco. We can see that for different Manhattan neighborhoods we get different results. However, some neighborhods in San Francisco turned out to be similar to more Manhatten neighborhoods than others, especially Park and Tenderloin

Conclusion

If there is a specific neighborhood in Manhattan that the customer wants to find similar neighborhood our tool could provide it. On the other hand, if the customer wants to know which neighborhoods are most similar to Manhatten area in general, Park and Tenderloin are valid options.