Neighborhoods comparison between cities using Data Sciences

Sudhir Chamarthi

September 25, 2019

1. Introduction / Business Problem

1.1 Background

The Business is a Real Estate Company specializing in relocations that assists clients moving from one city to another. Along with helping the customer find a property at their desired price, they would also like to provide a service of helping them choose a neighborhood based on social venues such as Cafes, Museums, Restaurants and Parks. Most customers place a premium on their quality of life and the social venues in a neighborhood contribute significantly towards that measure. This service would also try and let customers find neighborhoods in their new city that are comparable to the neighborhoods that they liked in the city that they are moving from. The business is very keen to implement this service to differentiate themselves from others in the market and has identified it as a key business objective.

1.2 Problem

Cities and Neighborhoods vary significantly and there are a very large number and types of social venues. Trying to collate such a large amount of data and finding meaningful relationships and comparisons between neighborhoods within a city or neighborhoods between cities is a daunting problem. A lot of the data about the Social Venues is also descriptive and so data analysis is further complicated.

1.3 Interest

Anyone interested in buying/renting a property in a city or anyone moving from one city to another and looking to search for their ideal neighborhood

to live-in would greatly benefit from this service. The service provides insights into the social makeup of their potential neighborhood. Further, it provides them with an opportunity to compare neighborhoods between cities, which will have a bearing on their quality of life measures.

2. Data

2.1 Data Sources

The three primary data sources we will be using for the solution we are attempting are Wikipedia, GeoPy Python Library and FourSquare API.

Wikipedia is a freely available website that is user curated and monitored and is well accepted as a good source of information. We will be using the text from the wikipedia pages to scrape the names of the neighborhoods. In this particular exercise we will use the neighborhood listings of the two cities we are comparing, New York City and San Francisco. The data sources of neighborhoods are at the following URLs:

https://en.wikipedia.org/wiki/List_of_neighborhoods_in_San_Francisco
https://en.wikipedia.org/wiki/Neighborhoods_in_New_York_City

Geopy is a Python library that provides the Latitude and Longitude given a text address. It is freely available from https://geopy.readthedocs.io/

FourSquare Inc. is a Social Platform app and service that has vast trove of information about cities, especially pertaining to social venues such as Cafes, Museums, Restaurants, Parks and such. It lets users write reviews and recommendations and also provides an API to access their data set. We will be using their free basic API to pull down information about various venues/businesses in a neighborhood. The API requires the Latitude and Longitude of a neighborhood which will then be used to pull down the list of venues. The API returns the list of venues in a JSON format which is a popular format for interchange of text data on the internet.

The Neighborhood Data from Wikipedia and the Social Venues data from Foursquare along with the Geo-location data from GeoPy Library will be

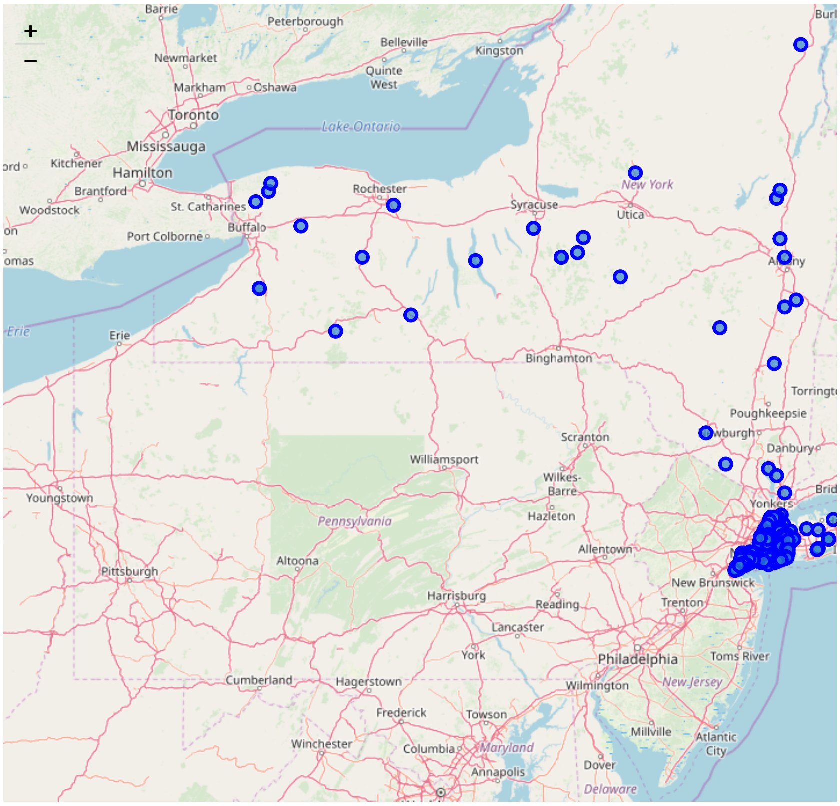
combined using an unsupervised Machine Learning k-means Clustering algorithm to find groups of neighborhoods in the two cities that are alike. The clustering algorithm is well suited to try and group a large number of data points to desired number of clusters. The data from the above sources suffices for our purposes. The data gathered is cleaned, transformed, manipulated and visualized using a few Data Science libraries such as pandas, numpy, Matplotlib before being finally fed to the Clustering algorithm to get the desired outcome of groupings of neighborhoods between cities. A caveat is that if the two cities vary considerably in their Social Venue makeup, then the clustering results will not yield good groupings and this is acceptable.

2.2 Data Cleaning

After obtaining the Latitude and Longitude of the neighborhoods and visualizing the neighborhoods with a map, we observed that certain neighborhoods were not located in the City that was being mapped. This was due to the GeoPy library not successfully mapping the Latitude and Longitude the address provided. Since we are trying to reconcile two different data sets it isn't surprising that there is a mismatch. Perhaps the library expected a more precise textual address to be provided. For our purposes, since the number of mismatches are a small percentage we have decided to remove the mismatches by identifying the statistical outliers. We used the Latitude and Longitude numbers to calculate outliers to remove neighborhoods that were below Q1 - 1.5*IQR and above Q3 + 1.5*IQR where Q1 and Q3 are the first and third quartiles and IQR is the interquartile-range.

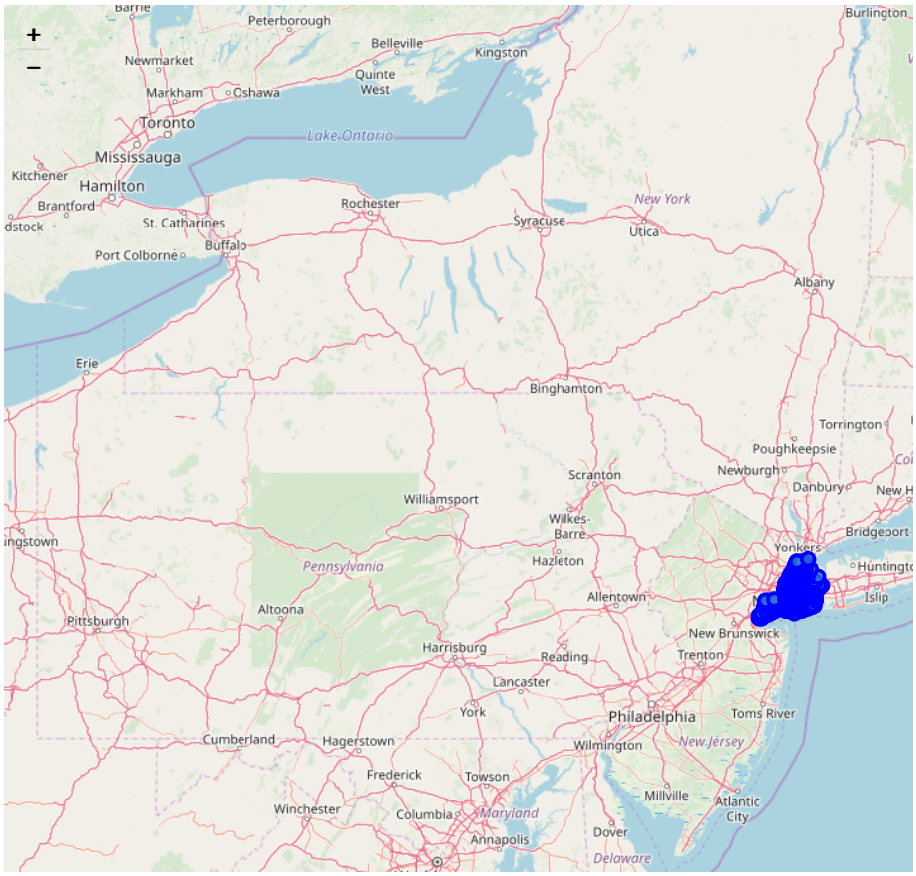
Further, a river that was wrongly identified on the Wiki as a neighborhood ('Bronx River') was removed.

Before clean-up:



Leaflet (https://leafletjs.com) | Data by © OpenStreetMap (http://openstreetmap.org), under ODbL (http://www.openstreetmap.org/copyright).

After clean-up:



data:text/html;charset=utf-8;base64,PCFET0NUWVBFIGh0bWw+CjxoZ...NjM1Y2M0MWIpCiAgICAgICAgOwoKICAgICAgICAKICAgIAo8L3NjcmlwdD4=

From the FourSquare API Venue data, “Neighborhood” was listed as one of the Venue Categories. Since we are interested in venues in a neighborhood, entries with the “Neighborhood’ venue category label were removed.

2.3 Data Transformation

Much of the data we will be using for the project is ‘Categorical’ except for latitude and longitude which are quantitative. The categorical data is transformed into a quantitative data using ‘One hot encoding’ technique to help with the final data modeling and analysis. The encoding assigns a numerical number for each type of Categorical data.

2.4 Data Selection

Using the FourSquare API we were able to get Name, Latitude, Longitude and Category for all the venues of the neighborhoods in the selected Cities. We further refined our data selection to just the venue categories since we are just interested in the top categories of Social Venues in a given City and Neighborhood. The same data selection also suffices for our Data Clustering objective of trying to group neighborhoods in cities based on the Social Venue categories. There were a total of 473 Venue Categories that we chose as our Data set on which to run our Data Clustering.