Analyzing Cities: A Case Study of Singapore with comparisons between Singapore and New York

Introduction/Business Problem

As cities' around the world grow larger it becomes increasingly important to understand their internal structure as well as the similarities and differences between cities. Focusing on Singapore I am using unsupervised machine learning techniques to identify clusters of similar neighborhoods within and across the five regions of Singapore (Central, North, North-East, East, and West). For this analysis I am defining similarity in terms of the types and number of amenities in the neighborhood. Thus, I describe each neighborhood as a vector of the amenities within the neighborhood. I augment this by demographic data on the residents of each neighborhood to gain a better understanding of which demographic groups reside most commonly in which kind of neighborhood. I will investigate whether similarity in terms of amenities correlates with similarity in terms of the demographic characteristics of residents.

Identifying neighborhoods that are similar to each other in terms of amenities and which demographic groups typically reside in them can help urban planners/city planners to make more informed decisions. It may also help the proprietors of restaurants, bars, etc. to make informed decisions about the location of new venues. Additionally, the results can help residents that need to relocate from one part of the city to another to find neighborhoods that are similar to the one they currently live in or to select a neighborhood with similar amenities.

Next, I create a similar dataset of amenities within neighborhoods in New York City and identify similar neighborhoods within and across boroughs of New York City. This part of the analysis mirrors the analysis of Singapore but excludes demographic data.

Based on the clusters of similar neighborhoods in New York and Singapore I will examine similarities across the two cities. This sort of across city comparisons can help, for example, expatriates who move from one city to the other identify neighborhoods that are similar to the one they currently live in.

Data

For this project I am combining several datasets. First, I am obtaining lists of neighborhoods and their geographical coordinates for Singapore and New York City.

Foursquare data

For both cities Singapore and New York City, I obtain data on venues located within the neighborhoods/subzones. I will use the Foursquare API to obtain the venues data. Specifically, for each city and each neighborhood/subzone I search venues within a certain radius around the neighborhood/subzone center. For each venue I extract the venue category to create a profile based on the frequency of venue categories within each neighborhood/subzones. The frequencies of venue categories serve as input for unsupervised machine learning algorithms.

New York City Location Data

For New York City I am using the data provided in the IBM lab under the link https://cocl.us/new_york_dataset (alternatively via https://geo.nyu.edu/catalog/nyu_2451_34572)

The datasets provide names and latitude, longitude coordinates of neighborhoods in New York City as well as the name the borough the neighborhood belongs to. This data allows me to search for venues within each neighborhood using the Foursquare API.

Singapore Location Data

For Singapore I am using data made available by the government of Singapore. Singapore's government provides a shapefile that contains the boundaries and centers of the so-called subzone of the city-state. For each subzone the dataset also contains the name of the region the subzone belongs to. The file can be freely downloaded via https://geo.data.gov.sg/mp14-subzone-web-pl/2014/12/05/shp/mp14-subzone-web-pl.zip

Using the definition of the Singaporean government, these subzones are defined as follows: *"Subzones are divisions within a planning area which are usually centered around a focal point such as neighborhood center or activity node."* There are in total 323 such subzone across the 5 regions of Singapore. Most subzones (139) are located in the central region. For this analysis, I will treat Singapore's subzones as equivalent to New York City's neighborhoods. Similar to New York City this data allows me to search for venues within each of Singapore's subzones using the Foursquare API.

Singapore Demographics Data

In addition to the subzone locations I obtain some socio-demographic data for each subzone which is also conveniently provided by the Singaporean government. The socio-demographic data is provided via https://data.gov.sg/dataset/resident-population-by-planning-area-subzone-age-group-and-sex-2015?resource_id=68775b41-3025-4763-970d-f479652e8b05 and https://data.gov.sg/dataset/resident-population-by-planning-area-subzone-and-type-of-dwelling-2015?resource_id=2719aca1-6b37-4f8b-ac3f-a866d32df7c6

The socio-demographic datasets represent the results of the 2015 wave of the Singapore household survey. For each subzone the socio-demographic datasets include the total resident population and breakdowns of the resident population by age group, type of dwelling, and ethnicity. This data allows us to analyze the similarity between neighborhoods of Singapore in terms of the demographic characteristics of its residents, enabling the comparison between neighborhood similarity in terms of amenities and terms of demographics.