Introduction

A description of the problem and a discussion of the background

Let's assume that one has job offers from a company located in Seoul, S.Korea, he stays in Manhattan, New York City at the moment and want to find as similiar environment as possible to that of Manhattan.

- First, he want to segment & cluster the neihborhoods of the Seoul, S.Korea
- and determine which district in Seoul is most similar or dissimilar to Manhattan, NY.

Brief information about both cities: Seoul officially the Seoul Special City, is the capital and largest metropolis of South Korea. With surrounding Incheon metropolis and Gyeonggi province, Seoul forms the heart of the Seoul Capital Area. It covers an area of 16,000/km2 (42,000/sq mi) and has It has an estimated population of 9,838,892 as of 2018(https://en.wikipedia.org/wiki/Seoul)

Data

A description of the data and how it will be used to solve the problem.

The Districts (Gu) of Seoul are the twenty-five gu ("districts"; 구; 區) comprising Seoul, South Korea. The gu vary greatly in area (from 10 to 47 km²) and population (from less than 140,000 to 630,000). Songpa is the most populated, while Seocho has the largest area. Gu are similar to London's or New York's boroughs or Tokyo's 23 special wards, and a gu's government handles many of the functions that are handled by city governments in other jurisdictions. This city-like standing is underscored by the fact that each gu has its own legislative council, mayor and sister cities. Each gu is further divided into dong or neighborhoods. Some gu have only a few dong while others (like Jongno-gu) have a very large number of distinct neighborhoods(https://en.wikipedia.org/wiki/List_of_districts_of_Seoul)

- Seoul.csv

To get longitude and latitude information for each area, get data from a Korean govenment site(https://www.data.go.kr/dataset/3045281/fileData.do). There are 25 files that include addresses and coordinates of each area of 25 districts in Seoul. 1 file is erroneous though. So I had to manually insert coordinate information of coordinates of 'Dojak-gu' district. Each coordinate of area is calculated by mean() function of every addressable building coordinates' information of the area. For the convenience, I downloaded those 24 files, merged them into 1 file, and placed .csv file on the server, so you can simply use it.

- Seoul_coord.csv

Using coodinates information, we'll use k-means clustering method of machine learning.