Relocation Project

Submission for Coursera Capstone Project

Introduction

Relocation is the action of moving to a different place and establishing a residence there. Relocation is done by people for many reasons such as for a new company position in a different place or for marriage reasons. This project aims to provide information for people looking to relocate between the districts in Bangkok and the districts in Chiang Mai, both of which are provinces with large population in Thailand. Specifically, this project will provide groups of similar districts in the sample.

Target Audience

The target audience of this project is for those who need to relocate from Bangkok to Chiang Mai or vice versa. There are various reasons for relocation including moving to accept a new company position at a different location or moving for marriage reasons. Those people will benefit from this project because they can then negotiate or figure out the best course of action for relocating.

Scope

The scope of this project is limited to Bangkok and Chiang Mai only. This is to limit the requests made on the FourSquare API. Furthermore, this project will only use features related to lifestyle. This includes the types of venues around a particular area and the density of people in that area. These limitations are made to keep the project completable under 2 weeks as required by Coursera.

Data

As this project aims to get provide the groupings of districts in Chiang Mai and Bangkok from lifestyle aspects, this project will be looking at the proportion of venue types in each districts. Additionally, the population density of each districts will also be taken account.

Proportion of venue types

The venue types located in each districts is taken from FourSquare API. The /venues/explore endpoint is used to retrieve the list of venues along with its name, latitude, longitude, and

category. The category as obtained from this endpoint is very specific. It can be generalized by using the /venues/categories endpoint which contains the whole hierarchy of available categories. The proportion of venue types is then calculated by the number of a specific venue type divided by the total number of venues in that area.

The proportion of venue types is important because this number broadly shows the specialization of that district. If it contains more business / professional venues, then it is likely to be some business center. If it contains a lot of hotels then the district's specialization might be toward tourism.

Population Density

The population density is calculated by the district population divided by the district area. The district population data is taken from the Bureau of Registration Administration (BORA) and the district area is taken from the Energy Policy and Planning Office (EPPO). The district population data consists of the male population, the female population, and the total population, available annually from 2000 to 2017. The district area data consists of only the district area in squared kilometers.

The population density represents the district in many ways. The population density can be used to estimate the capacity of the venues in that district and also approximate the amount of social activities in that area.

Methodology

Firstly, the list of district names that are located in Bangkok and Chiang Mai is needed. This information is scraped from a website that lists both the Thai name and English name. Bangkok currently has 50 districts while Chiang Mai currently has 25 districts. The list of district names is then collected into a Dataframe, a data format that is easy to manipulate and perform further analysis. This Dataframe is called districts_df and will be preserved throughout the process.

Secondly, the latitude and longitude for each districts is fetched. This is accomplished using the Nominatim agent with a fallback to the Photon agent in case the Nominatim agent couldn't find the latitude and longitude for the specified district. The Photon agent was fell back twice during the process; once for 'Mae On, Chiang mai, Thailand' and another for 'Galyani Wattana, Chiang mai, Thailand'.

Upon plotting the fetched coordinates on a geographical map, it seems that 'Galyani Wattana, Chiang mai, Thailand' is significantly far away from Chiang Mai and 'Phra Nakhon, Bangkok, Thailand' is also significantly far away from Bangkok. To circumvent this, 'Galyani Wattana, Chiang

mai, Thailand' is dropped and 'Phra Nakhon, Bangkok, Thailand' is fetched again using the Photon agent. This leaves Chiang Mai with only 24 districts in the system.

Next, the population data is scraped. This information is scraped from a website that shows the female population, male population and total population for every districts in Thailand. The website provide population data from 2000 to 2015, annually. However, the subdivisions for Chiang Mai were partially different from the subdivisions one would typically see; the subdistricts were listed among the districts. A manual search on the internet need to be made in order to group the subdistricts back into the districts they belong to.

The district area data is scraped after that. The area data is listed along with the the districts names for this webpage. There were some problems while scraping the list of districts for Bangkok site; the list contains both districts and subdistricts in the same page. Fortunately, the names were prefixed to show their type. Another problem is that this website is using some Microsoft Word Framework to generate their website and the data can't be downloaded without using a more sophisticated method of scraping. To keep the complexity to a minimum, the webpage is download and stored locally.

The population density is now ready to be calculated. The calculation is made by dividing the population data by the area for each district. Two years were selected: 2012 and 2017. There two values were selected because 2017 is the most recent year and 2012 is significantly old enough to provide statistical trend about the population.

Next, the venues data is fetched from the FourSquare API. The data is taken from the '/venues/explore' endpoint. This endpoint is fetched for each district multiple times until the all the venues in the area have been fetched. The category is also mapped to its top level category because the number of category is very large and there may be too many features to effectively perform clustering.

The features that will be used for clustering is then selected. The chosen contains the population density for 2012, the population density for 2017 and the ratio of each venue category in each district. To standardize the values, all the values were fed into a standard scaler to normalize the values so that clustering would be more effective.

Clustering is performed using the KMeans algorithm. The number of clusters is chosen to be 5 as this number deemed fit for the task at hand. The clustering is then plotted on a geographical map to show the distributions of the groups, the similar between districts, and to verify the validity of the groupings.

Results

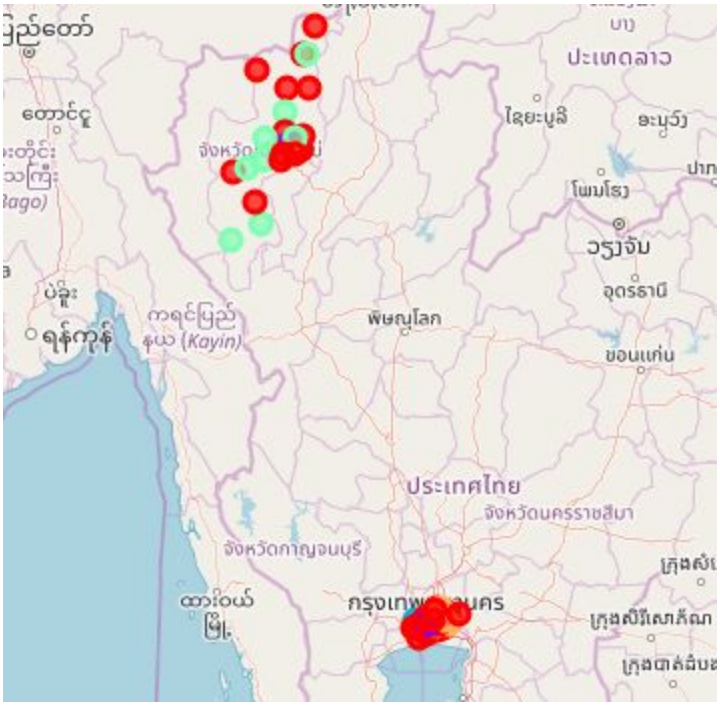

Fig 1 - An overview map showing both Bangkok and Chiang Mai

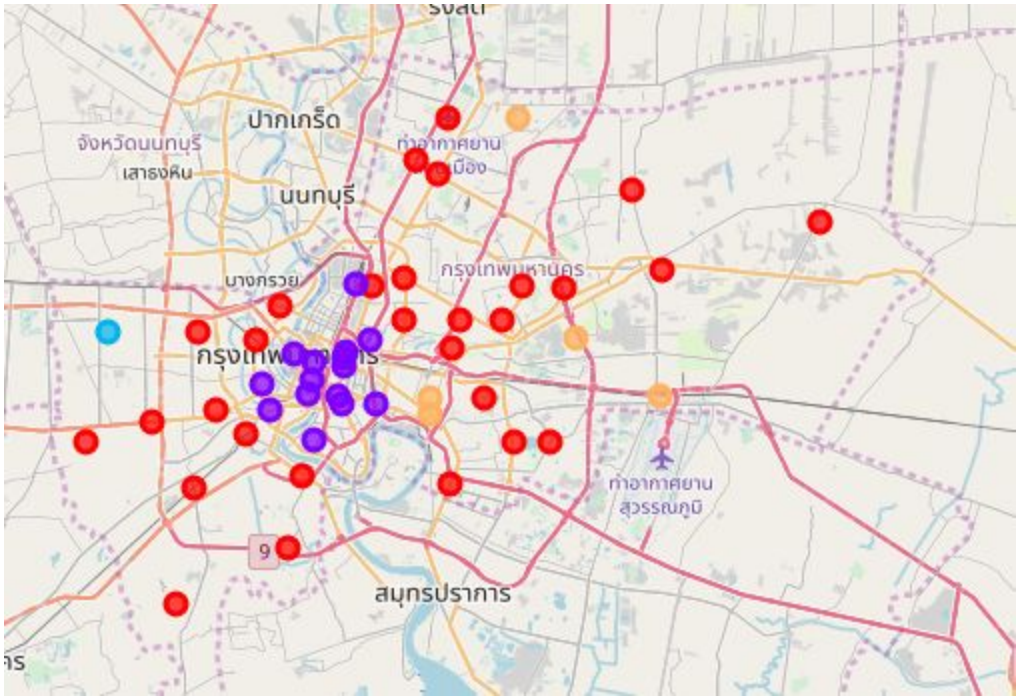

Fig 2 - A map showing the clustering for all 50 districts in Bangkok, Thailand.

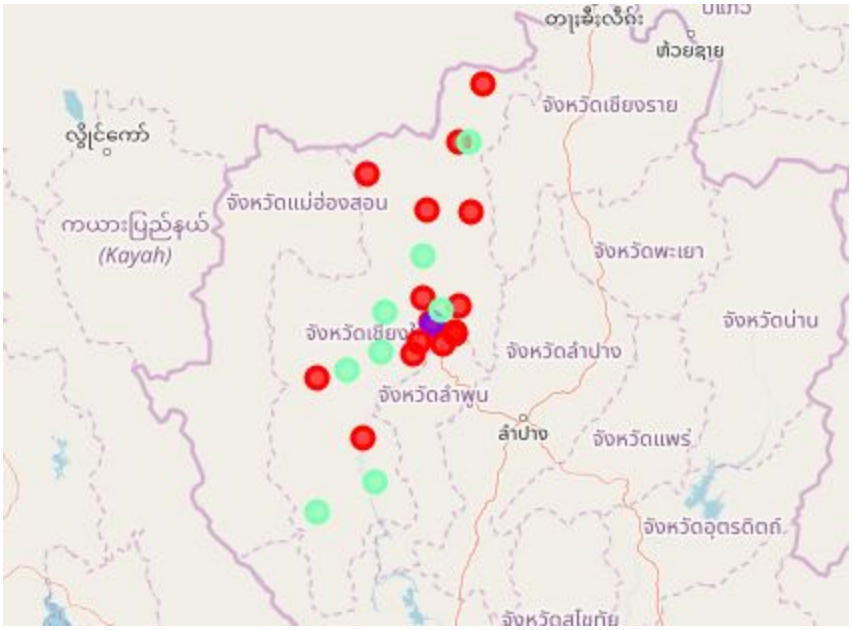

Fig 3 - A map showing the clustering for 24 districts in Chiang Mai, Thailand.

Discussion

It seems that there are indeed districts in Bangkok that are similar to Chiang Mai. Some relocations are more probable than others; the suburbs of Bangkok is very similar to the city area of Chiang Mai while the outskirts of Bangkok are similar to the suburbs of Chiang Mai. The city center of Bangkok is unlike any area in Chiang Mai except for Mueang Chiang Mai. On the other hand, the outskirts of Chiang Mai is also very different from all districts in Bangkok; therefore, a lower possibility exists for relocation for those in either the city center of Bangkok or the outskirts of Chiang Mai.

Conclusion

This project has used data as obtained from the FourSquare API as well as other sources to cluster the districts based on similar lifestyles as determined from population density and types of venues in the area. People who are interested will be able to use this data to make better decisions regarding relocation between districts in Bangkok and Chiang Mai.