Similar Neighborhoods in Finland

Hannu Niittymaa

April 5th 2019

1 Business Problem

This chapter describes what is the background and actual problem that this data science project is intended to solve.

1.1 Description of the problem

In Finland and overall in the world, an urbanization is the strengthening trend. People are moving from countryside and smaller cities to bigger and thus more vital cities. Also, people are moving between cities due to different reasons, like getting a new job or studying place. Moving to the new city is not a minor task and it would help a lot to get some information beforehand on neighborhoods' characteristics and how they compare to some known cities neighborhoods.

In this data science project, I try to find out, what are the similarities and differences between neighborhoods in the tenth biggest city in Finland. This is done by investigating the similarity or difference of cities in terms of the number of venues and the type of venues. As some cities are quite a big and neighborhoods differ from each other's within city, the study is done at neighborhood level based on postal code. The neighborhoods are clustered to five clusters and each neighborhood is shown on map to make it easy to compare neighborhoods within and between citites.

1.2 Interest

There are many groups that are interested in the results of this project. For example, unemployed people, who cannot find employment from their current home city or neighborhood. In case that they are looking for new job from other cities, they can use the results to think about appealing and suitable cities for them. Students, who are considering where to study, can also use results of this project. For example, in Finland it is possible to study economic science at least in six different university in different cities. The entrance examination is common for all universities and student has to select the order of their most favorite universities before entrance exam. This project's results help students to compare the appeal of different cities and put the universities in order.
In general, all people who are considering moving, can use the results to consider potential new living place.

2 Data

2.1 Data sources

Statistics Finland, a national statistical authority in Finland founded in 1865, provides lot of information of cities and their neighborhoods including postal codes. It produces the vast majority of Finnish official statistics and is a significant international actor in the field of statistics. Lot of their data is open data and available for everyone. The actual data needed in this project is the ten biggest cities based

on the number of inhabitants and the neighborhoods of cities based on postal codes. All this data can be retrieved from the open data store of Statistics Finland.

The Foursquare API is used to get different venues and number of venues for each neighborhood in cities. The Foursquare API is available to everyone and free to certain limits.

2.2 How data is used

The number of cities to be researched will be limited to ten biggest cities in Finland. Data containing city information must be filtered to include only ten biggest cities. As the data is in excel format, there are addition rows and titles and other unneeded cells. So, the created data frame must be cleaned and formulated so, that only the relevant information is left to data frame. Same cleansing and formulating must be done also for postal code data.

These two sources of data are downloaded and cleansed separately. After that they must be combined into one data frame and the venues data must also be combine into this data. After that it is possible to cluster the data and show it in the map

As the study is done for each neighborhood based on postal code, the radius from the centrum of postal code location will be decided when some real investigation of suitable values is done during the actual project.