

# The Most Appealing Neighborhoods in Finland

Hannu Niittymaa

May 3rd 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 cities' neighborhoods that people already are familiar with.

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, which makes it easy to compare neighborhoods within and between cities.

### **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**

This chapter describes the data sources, which are used in this project and how data needs to be formatted before it can be used in analysis.

## 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. Statistics Finland 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. So, there will be two data sources, both in Excel-format:

1. Finland\_cities.xlsx
2. Neighborhoods\_by\_postal\_code.xlsx

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 Data formatting

The number of cities to be researched will be limited to ten biggest cities in Finland, so data containing city information must be filtered to include only ten biggest cities. As the data is in excel format, there are plenty of 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. Some cleansing is already done, when the excel-file is loaded into data frame. More data manipulation, like dropping and renaming columns, can be done within data frame.

These two sources of data are downloaded and cleansed separately. After that they must be combined into one data frame and then venues data will be combined into neighborhood data. Before venue data can be combined into neighborhood data, the coordinates of neighborhoods are needed. They will be retrieved using geopy library's geocoders method. It seems to very slow and sometimes very unreliable, so the better way would be, if there were neighborhoods' coordinates available somewhere. Unfortunately, I wasn't able to find any source for coordinates, so I have to rely on geopy's geocoder.

When coordinates and venues have been added to each neighborhood, it is possible to cluster the data and show it on 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.

## 3 Methodology

This chapter describes what machine learnings I used and why, exploratory data analysis that I did, and inferential statistical testing that I performed.

### 3.1 Machine learning algorithms

As I am comparing the similarity of neighborhoods, the most convenient way is to cluster neighborhoods based on the number of similar venues they have. There are several algorithms for clustering with pros

and cons. I chose K-Means clustering algorithm, because it is easy to understand and implement. It is also very fast, even though in this exercise the amount of the data is relatively small.

On the other hand, K-Means algorithm has also some disadvantages. First of all, you have to select the number of clusters by yourself. This is not always obvious, and it would be better, if the algorithm itself would suggest the ideal number of clusters, which is very useful insight of data as such. K-Means cluster also selects initial cluster centers randomly, which can yield to different results, when running K-means algorithm several times.

### 3.2 Exploratory data analysis


The “moving” parameters in this exercise are the number of clusters, the number of venues per neighborhood and the radius of neighborhoods when retrieving venues.

The number of clusters should not be too high, because it gives too many options. Now neighborhoods are clustered into five clusters. We can easily notice from results that most of neighborhoods belong to cluster types one, two or five. There are only couple of neighborhoods, which belong to cluster 3 or 4.

The number of venues per neighborhood included in analysis is now limited to ten. Ten venues seems to be adequate level as some venues seem to be quite irrelevant like juice bar.

## 4 Results

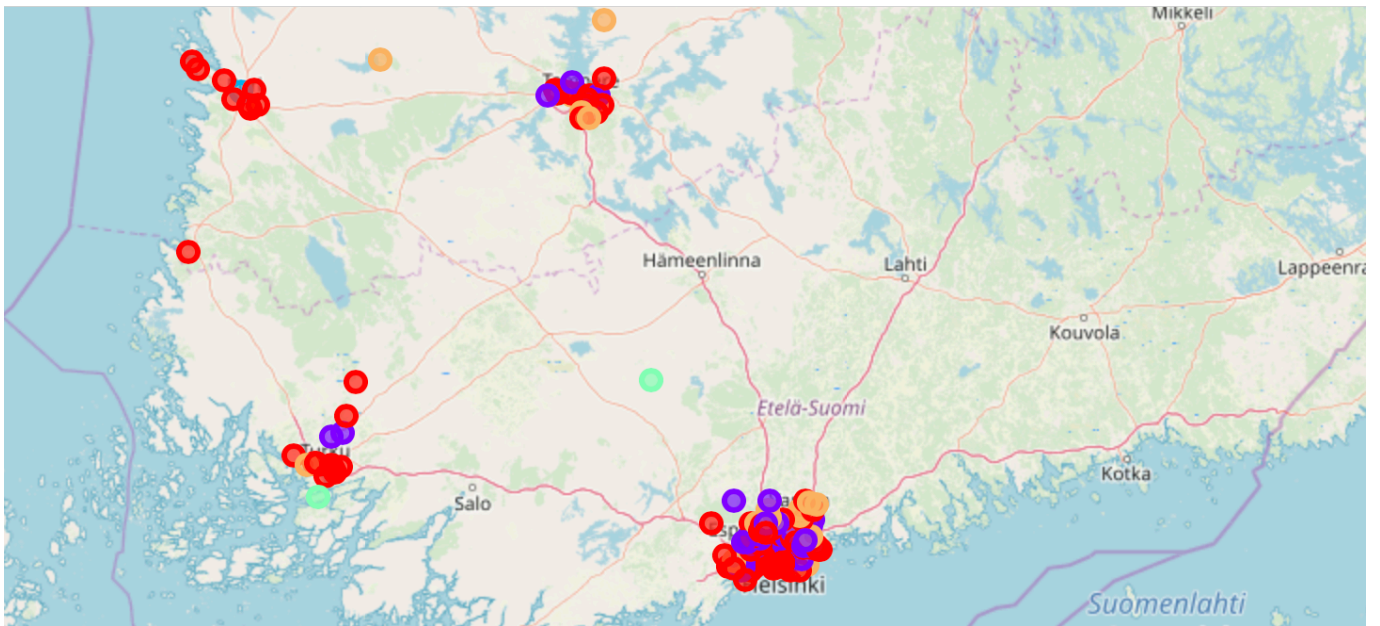
Ten biggest cities (by inhabitants) are listed below:



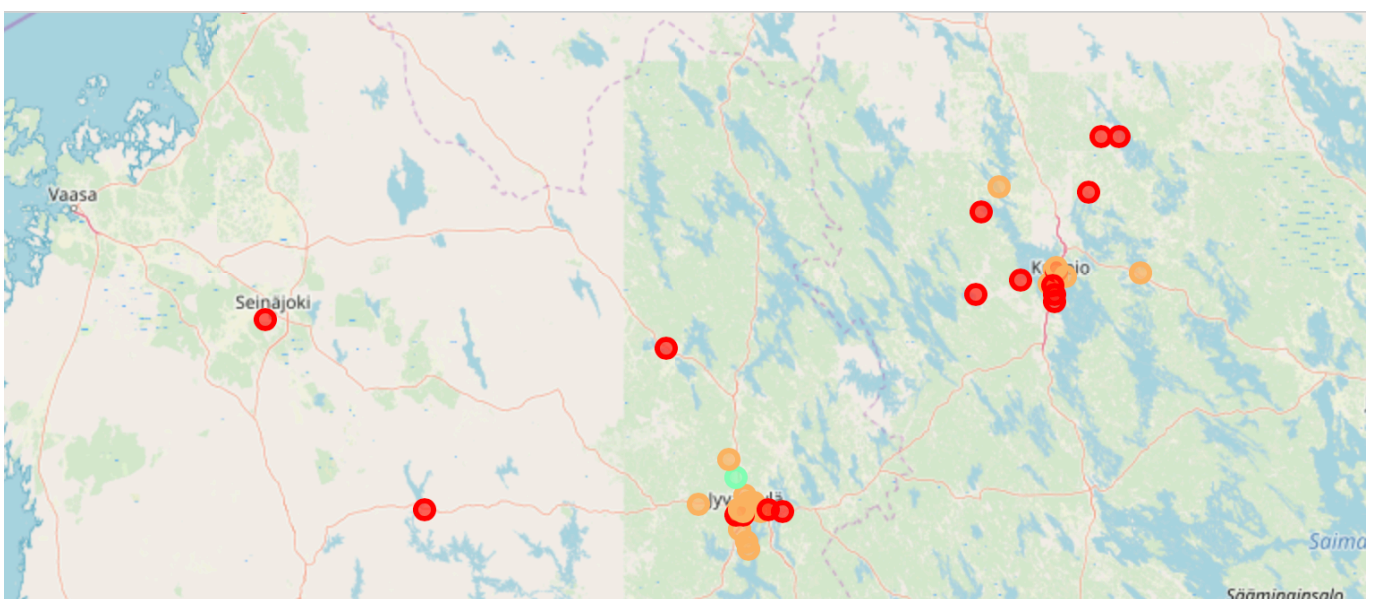
	City	Inhabitants
36	Helsinki	643272
14	Espoo	279044
271	Tampere	231853
294	Vantaa	223027
183	Oulu	201810
279	Turku	189669
66	Jyväskylä	140188
126	Lahti	119573
117	Kuopio	118209
204	Pori	84587

In the figures below we can see all clusters in each city.

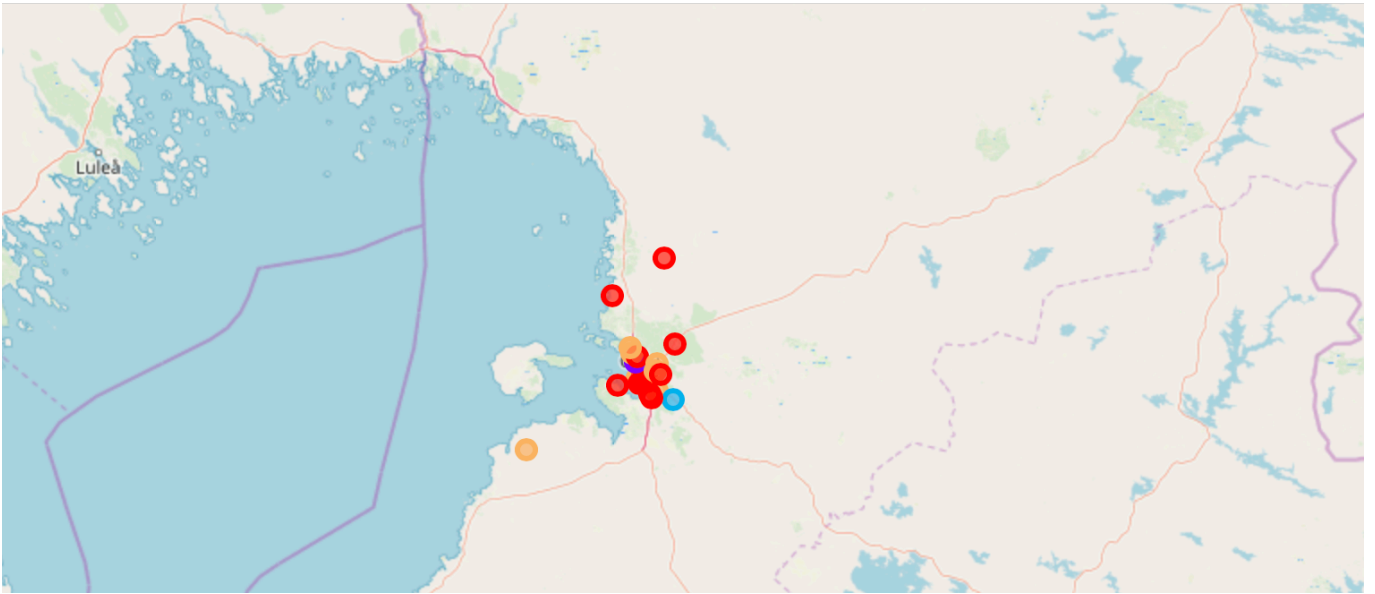
The Southern part of Finland including capital area cities Helsinki, Espoo, Vantaa and Tampere, Turku and Pori.



The middle part of Finland including cities of Kuopio and Jyväskylä.



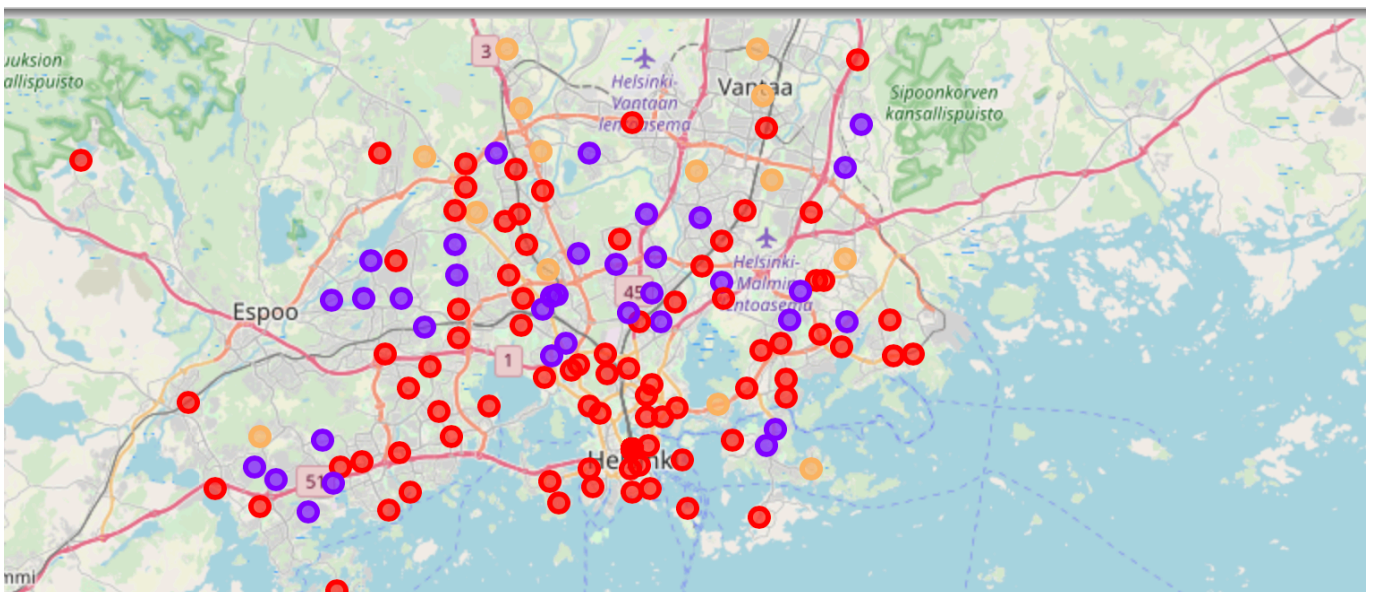
And finally the Northern part of Finland including city of Oulu.



We can easily notice that most of neighborhoods belong to cluster 1 (red), cluster 2 (purple) or cluster 5 (orange). There are only couple of neighborhoods, which belong to cluster 3 (blue) or 4 (green).

If we compare the similarity between cities, the quite obvious result is that in bigger cities there are more different neighborhoods than in smaller cities. In smaller cities there seem to be only cluster type 1 and 5 type of neighborhoods. In bigger cities there are also cluster 3 types of neighborhoods.

If we compare neighborhoods within cities or capital area as a whole (figure below), we can notice that cluster 1 (red) type neighborhoods mostly exists in the center of towns and then cluster 3 (purple) type neighborhoods and cluster 5 (orange) type neighborhoods exist outer part of cities.



By examining venues of each cluster, it is however not obvious, why neighborhoods are clustered that way.

## **5 Discussion**

I think that the biggest question in this study is that can we rely on the results? At first sight it looks like that there not so much in common between neighborhoods, which belong to same cluster (despite clusters 3 and 4). Also, the most common venue between neighborhoods within one cluster seems not to be so relevant, like bus stop in cluster 2.

In overall, this study would require more work and for example irrelevant venues should be filtered out from results. Also, adding also other data than venue information, for example combining each neighborhoods salary level into results, would bring much more insight when comparing neighborhoods. Salary level data is even available (provided by Statistic Finland), so it could be used.

Anyway, the current level of study fulfills perfectly the purpose of this study and gives a good understanding of data science project.

## **6 Conclusion**

In this study, I analyzed the similar neighborhoods in ten biggest cities in Finland. Neighborhoods were determined based on postal code and similarity between neighborhoods were determined based on the similar venues that exist in neighborhoods. Even though this study took a lot of time, it was a bit surprise how easy at the end it was to find necessary data and cluster that data using K-Means algorithm. I also learnt a lot during this exercise and this exercise certainly increased my motivation towards data science. I also hope that this study provides very useful tool for its interest groups.

It would be very interesting to expand this study to include other relevant data, like salary level, and also time dimension i.e. how neighborhoods in different cities have changed during past years. This combined with demographics data would provide very interest insight and expand this study's interest groups to demographers, city planners and architects, etc.