The Most Appealing Cities in Finland

3rd May 2019
Hannu Niittymaa

Introduction

- Data science project to find out similar neighborhoods in Finland.
  - Similar neigborhoods within city and between cities (ten biggest city in Finland).
  - Neighborhoods are defined based on areas' postal code.
  - Similarity is defined based on similar venues on each neighborhood.
- Background
  - Changing place of living is not a minor thing, so it would be nice to know better the characteristics of the new place of living beforehand.
  - Some cities are growing heavily and some regressing. This is a common trend and and it indicates the vitality of the city and changes can be quick. This study is one way to investigate the vitality of some city.
- Target audience
  - To all who are considering moving to new neighborhood within same city or between cities .
  - Students, job seekers, retired people etc.

Data Sources and Cleaning

- *Statistics Finland, a national statistical authority in Finland, founded in 1865, provides lot of information about cities and their neighborhoods including postal codes.*
  - Statistics Finland provides cities and related neighborhood data in Excel-format.
- Data cleaning and formatting
  - 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 on the data frame.
- Foursquare API
  - Venue data related to each neighborhood was retrieved thru Foursquare API.

Methodology used

- K-Means clustering algorithm was used to cluster the data.
- City, neighborhood and venue data needed to be combined into one data frame.
- Clustering was performed against combined data frame.
- Selected number of clusters was five.
- K-Means Pros
  - Easy to understand and implement.
  - Very fast
- K-Means Cons
  - The number of clusters must be determined by self, the algorithm itself does not suggest the optimal number of clusters .
  - Initial cluster centers are selected randomly -> results can be different on each algorithm run iteration.

Results

- Ten biggest city by inhabitants in Finland (on the right):

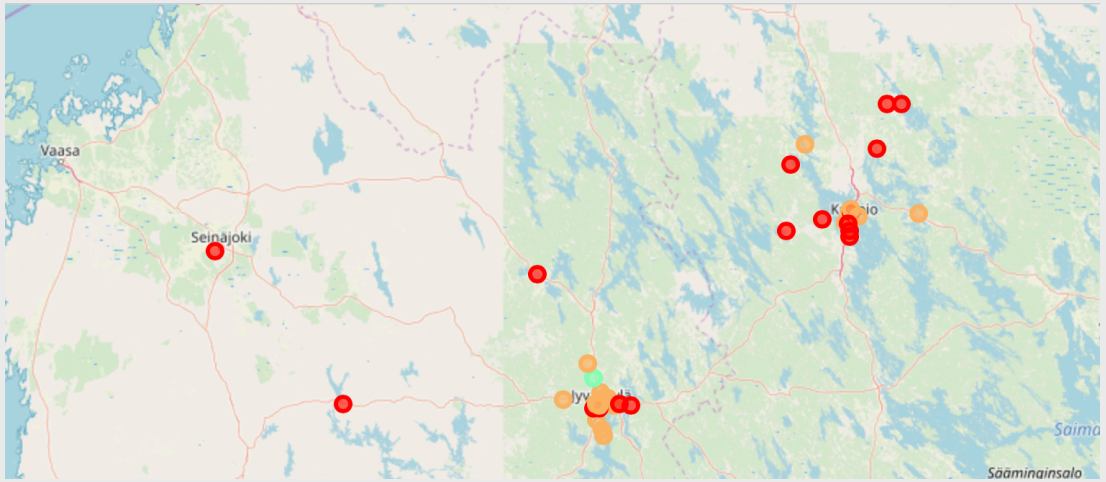
	City	Inhabitants
36	Helsinki	643272
14	Espoo	279044
271	Tampere	231853
294	Vantaa	223027
183	Oulu	201810
279	Turku	189669
66	Jyväskylä	140188
126	Lahti	119573
117	Kuopio	118209
204	Pori	84587

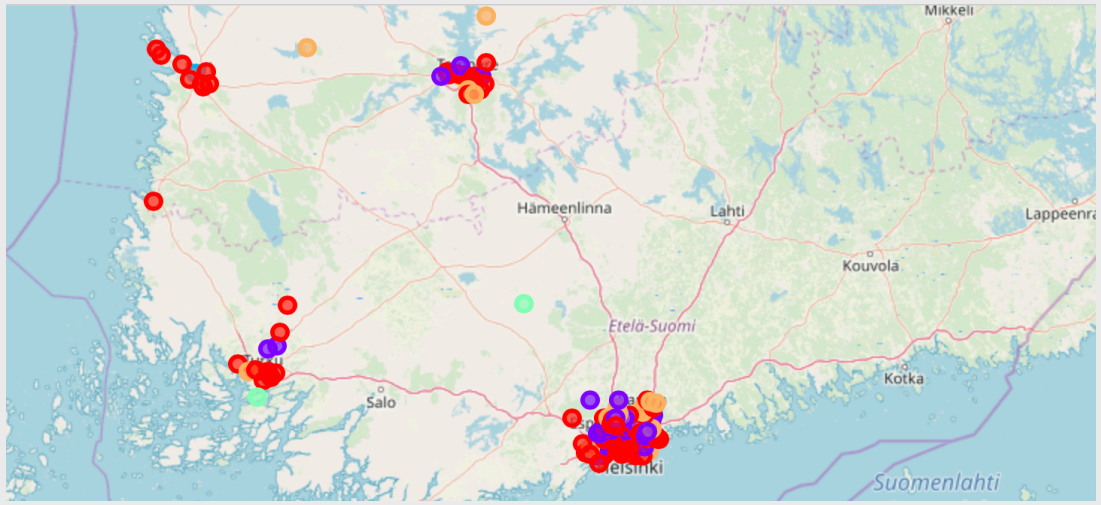
- The number of clusters were determined to be five (see next slides) and number of venues to be ten.
  - Cluster 1 (red), cluster 2 (purple), cluster 3 (blue), cluster 4 (green) and cluster 5 (orange)
  - Two of five clusters consisted only couple of neighborhoods (cluster 3 (blue) and cluster 4 (green))
- The summary of (obvious) results are that
  - in bigger cities there are more diversity between neighborhoods
  - In bigger cities there are ”neighborhood zones” around the center of city from venue point of view

Neigborhoods between cities

- The obvious result is that in the bigger city there are more different type of neighborhoods (cluster 1 (red), cluster 2 (purple) and cluster 5 (orange). This can be seen in lower picture on right hand side, where the capital area (Helsinki, Espoo and Vantaa) and Tampere and Turku are the biggest cities containing type 1,2 and 5 clusters.

- In smaller cities there are only cluster types 1 (red) and 5 (orange). This can be seen in upper picture on right hand side, where can be seen neighborhoods in Jyväskylä and Kuopio. Also in lower picture Pori (on upper left corner) consist only type 1 (red) clusters.

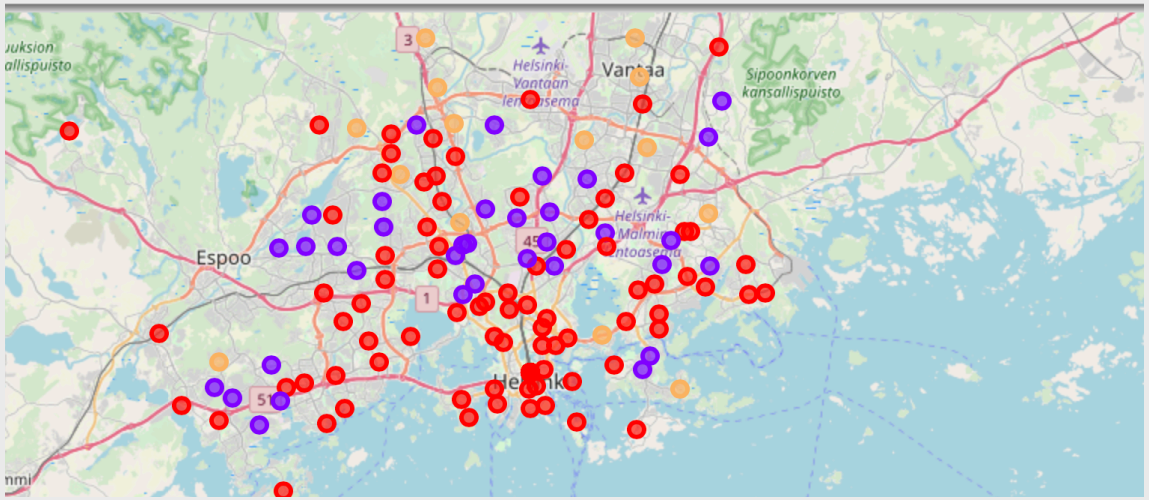




Neigborhoods within cities

- In picture on right hand side can be seen that in the capital area (Helsinki, Espoo and Vantaa) cluster 1 (red) type neighborhoods mostly exists in the center of the town and then cluster 3 type neighborhoods next and cluster 5 (orange) types of neighborhoods exists outer part of capital area.

- This same applies also, when investigating other bigger cities like Turku and Tampere.



Discussion and conclusion

- Are results reliable?
- Two of five clusters consisted only couple of neighborhoods-> was the number of clusters correctly defined.
- It was very difficult to percieve the differences between three other clusters.
- In one cluster the bus stop was the most common venue for most of the neighborhoods -> does this really distinguishes neighborhoods from each others.
- Most likely more data is needed to really distinguish neighborhoods from each others.
- For example salary data (provided by Statistics Finland) could be very valuable for further analysis.