The Battle of Neighborhoods

Helsinki (Finland) vs. Tallinn (Estonia)

Prepared by: Norbert Juhász

For: Applied Data Science Capstone by IBM/Coursera

Version date (Week 2): May 30th, 2020

1. Introduction/Study Problem

In this final course assignment, I will compare the neighborhoods of two cities:

- Helsinki, Finland (where I live) and
- Tallinn, Estonia (which is the closest other EU capital, just on the other side of the Gulf of Finland)

As a fact to raise the interest, there are even plans to connect the two cities by an underwater tunnel in the future. For more about this, please see https://en.wikipedia.org/wiki/Helsinki%E2%80%93Tallinn_Tunnel

Understanding the similarities of neighbourhoods between the two cities may help people who consider moving from one city to the other. For someone who, for instance, would be looking for a similar neighborhood to live in Tallinn as compared to where they live in Helsinki, the results of my analysis may be helpful.

In this project assignment, the neighborhoods of the two cities will be clustered, analyzed and described.

2. Data

Based on definition, my primary interest is profiling the neighourhoods of Helsinki and Tallinn with the purpose to cluster them based on similarities.

The following data sources are used:

- neighborhoods of Helsinki and Tallinn are defined based on postal codes from **Opendatasoft.com**'s API
    - dataset name: "geonames-postal-code%40public-us"
    - data of interest: postal code, area name, latitude, longitude, administrative structure
    - **Helsinki** is defined as Greater Helsinki area, including the administrative (but not geographically) separate cities of Espoo, Kauniainen and Vantaa
    - **Tallinn** city is also extended by the region of Viimsi vald which encompasses Tallinn towards the direction of Helsinki
- the central coordinates of Helsinki and Tallinn are retrieved from **geocoders Nominatim**
    - in order to center the folium map and make both cities visible, the coordinates of the two cities are averaged
- the venues, their type and location in every neighborhood will be obtained from the **Foursquare API** (as required by the assignment specifications)

3. Methodology

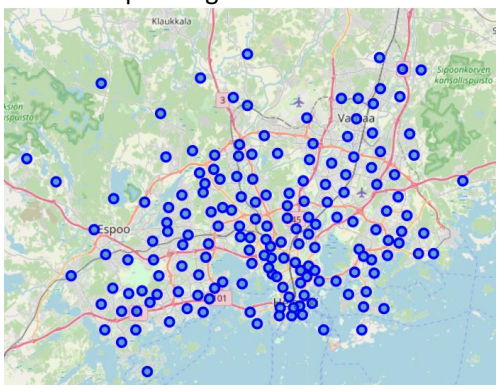
In this analysis, the focus is on clustering the neighborhoods of the Helsinki capital region and Tallinn based on the similarities of their venue profiles.

Firstly, the collected data is explored and visually analysed to ensure that it is suitable for the purpose.

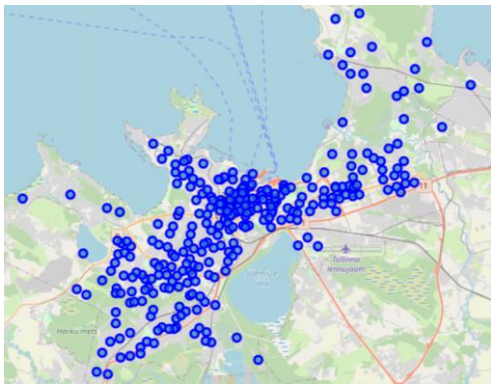
The loaded **postal code data** from Opendatasoft API has yielded 169 postal codes for the Helsinki capital region and 310 postal codes for Tallinn. Based on the population of the two cities (about 1.2 million for Helsinki and about 450 thousand for Tallinn), the expectation was to find more postal codes for Helsinki. However, clarifying the administrative differences between the two countries is out of scope from this project.

On the **folium map**, centered by coordinates retrieved from geocoders Nominatim, the neighbourhoods look visually correct and Tallinn has visibly more postal codes around the city center.

Helsinki capital region



Tallinn

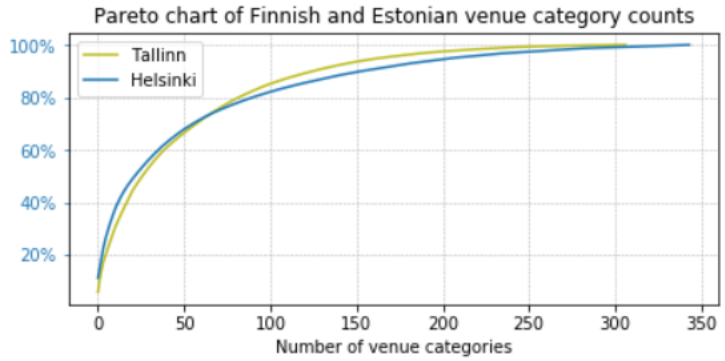


The loaded **location venue data** has yielded 17 538 records. This is a result of a Foursquare API request to explore the radius of 900m from each postal code coordinate and to retrieve a list of the maximum 100 most popular venues for each of them.

Checking the venue list of a well-known postal code area and browsing through it confirms that the extract was successful. The places are known already right on top of the list.

postal_code	latitude	longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
02600	60.2122	24.8066	Mezza	60.215323	24.811755	Middle Eastern Restaurant
02600	60.2122	24.8066	Jungle Juice Bar	60.218400	24.813196	Juice Bar
02600	60.2122	24.8066	Caffi	60.218603	24.812755	Coffee Shop
02600	60.2122	24.8066	Fressi	60.216913	24.818744	Gym / Fitness Center
02600	60.2122	24.8066	ELIXIA Sello	60.218130	24.812810	Gym / Fitness Center

The location venue data contains 417 unique venue categories. A Pareto analysis reveals that approximately 75-80 venue categories cover 80% of the total venue count. The Pareto curve is similar for both capital regions. It is also visible from the chart that there are unique categories for both Tallinn or Helsinki.



110 venue categories (26.4%) are unique for Helsinki and 73 venue categories are unique for Tallinn (17.5%). The two cities have a different historical and cultural profile. This explains and justifies the prevalence of unique venue categories. Having this in the data is seen as giving a positive flair to the clustering algorithm, allowing it to take into account possibly very different neighborhood profiles.

On the other side of the coin, 234 venue categories are shared between the two city regions (56.1%), representing a 90.5% share of the total venue count.

Unique venue categories for Helsinki capital region

country_code	EE	FI
Venue Category		
Himalayan Restaurant	NaN	38.0
Waterfront	NaN	17.0
Taxi Stand	NaN	17.0
Platform	NaN	15.0
Karaoke Bar	NaN	13.0
...	..	...
Molecular Gastronomy Restaurant	NaN	1.0
Outdoor Gym	NaN	1.0
Canal Lock	NaN	1.0
Carpet Store	NaN	1.0
Road	NaN	1.0

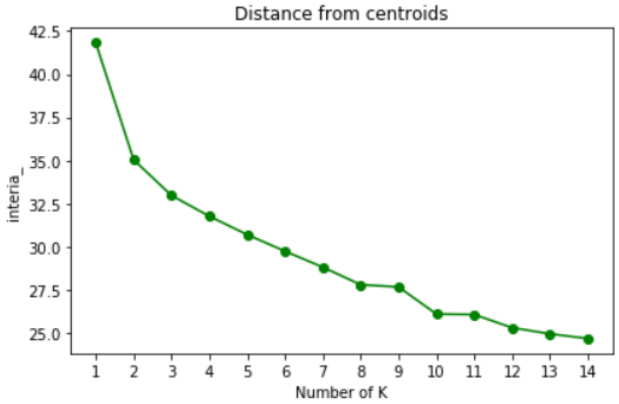
[110 rows x 2 columns]
Total venues in unique Finnish venue categories: 406

Unique venue categories for Tallinn

country_code	EE	FI
Venue Category		
Eastern European Restaurant	258.0	NaN
Bus Line	111.0	NaN
Market	59.0	NaN
Church	55.0	NaN
Shoe Store	53.0	NaN
...	...	..
Ski Chairlift	1.0	NaN
Residential Building (Apartment / Condo)	1.0	NaN
Fruit & Vegetable Store	1.0	NaN
Animal Shelter	1.0	NaN
Veterinarian	1.0	NaN

[73 rows x 2 columns]
Total venues in unique Estonian venue categories: 1265

Secondly, the **k-Means clustering algorhitm** is run in a loop in order to try various number of clusters (2 to 15) on the formatted dataset. The inertia is extracted from each model and plotted as visible on the below chart.

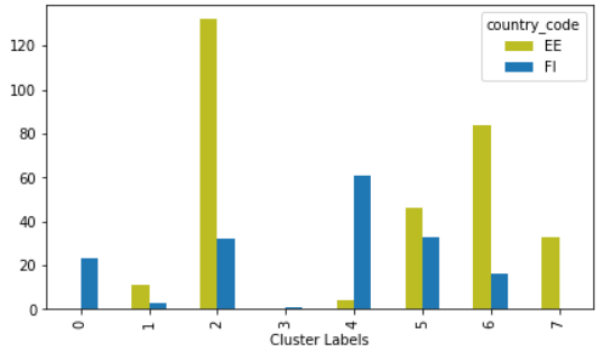


K = 2 was tried in line with the elbow method recommedation, but it was not found to be useful to identify similar clusters between the two cities. Therefore, **k= 8 was selected** as the next best option that would still prevent the model from overfitting.

4. Results

The k-Mean algorhitm has identified 7 clusters for Helsinki and 6 for Tallinn. There are only 5 clusters that overlap the two cities, with 2 clusters being unique for Helsinki and 1 for Tallinn.

The number of neighborhoods (postal codes) by cluster:



The clusters are described below and the top frequency venues are shown in the report. However, it needs to be mentioned that the k-Means clustering algorhitm took into account the full scope of avaialable venue categories, so there might be additional reasons for the cluster that are not directly visible from the top venue list.

Cluster 0 appears only in Helsinki (**red** colour on the map).

It seems to represent such remote neighborhoods where having a bus stop is a critical (and popular) part of the neighborhood profile.

This cluster also contains venue categories that represent sport and smaller stores to buy food.

Labels	Venue Category	Freq
0	Bus Stop	0.510887
0	Grocery Store	0.067734
0	Pizza Place	0.058091
0	Beach	0.019928
0	Soccer Field	0.018711
0	Playground	0.017475
0	Gym / Fitness Center	0.015259
0	Park	0.014612
0	Convenience Store	0.013686
0	Trail	0.011702

Cluster 1 is a small cluster overlapping both cities (**purple** on maps).

In both cities, it represents neighorhoods close to mostly the seashore, but also to nature and historical sites.

Labels	Venue Category	Freq
1	Beach	0.254762
1	Historic Site	0.107143
1	Harbor / Marina	0.079762
1	Boat or Ferry	0.053571
1	Bar	0.047619
1	History Museum	0.035714
1	Food Court	0.032143
1	Playground	0.026190
1	Campground	0.026190
1	Forest	0.023810

Cluster 2 clearly describes the city centres (**dark blue** on maps) and a few other central neighborhoods in other parts of the cities.

This is one of the biggest clusters and it is mainly described by the prevalence of popular venues, such as cafés, restaurants, parks, hotels, bars etc.

Labels	Venue Category	Freq
2	Café	0.062043
2	Restaurant	0.050510
2	Park	0.030481
2	Hotel	0.030082
2	Coffee Shop	0.024791
2	Pizza Place	0.016718
2	Eastern European Restaurant	0.016291
2	Bar	0.016153
2	Burger Joint	0.015776
2	Scenic Lookout	0.015748

Cluster 3 is the smallest cluster made up of only one postal code in Helsinki (**light blue** on maps), one construction & landscaping venue in the northern part of the city.

Labels	Venue Category	Freq
3	Construction & Landscaping	1.0

Cluster 4 is a cluster with many postal codes widespread all over Helsinki, but only a small number of postal codes in Tallinn in the south of the city (**cyan** on maps).

This cluster profile looks similar to Cluster 0, but the frequency of having a bus stop as a top venue is not as large as in Cluster 0.

This cluster would thus represent those nature intensive parts that still have good access to supermarkets, cafés etc. This can be also confirmed by own experience with having lived in such regions.

Labels	Venue Category	Freq
4	Bus Stop	0.264373
4	Grocery Store	0.051638
4	Park	0.041050
4	Pizza Place	0.039794
4	Soccer Field	0.039355
4	Supermarket	0.036313
4	Trail	0.021081
4	Playground	0.019473
4	Beach	0.018395
4	Café	0.018376

Cluster 5 seems to be yet an additional step away from Cluster 0 and Cluster 4 (**light green** on maps). The popularity of having a bus stop further decreases and the importance of other categories increases.

This cluster is well represented in both cities and it often borders Cluster 2 (city centre). In Helsinki, forest areas are widespread and are easily reachable from basically anywhere. This is not similar in Tallinn, so it can explain the shift of cluster count between the cities.

Labels	Venue Category	Freq
5	Bus Stop	0.092222
5	Park	0.077513
5	Pizza Place	0.038554
5	Café	0.029687
5	Supermarket	0.024467
5	Restaurant	0.023945
5	Trail	0.022948
5	Gym / Fitness Center	0.021179
5	Grocery Store	0.018181
5	Diner	0.016424

Cluster 6 is also represented in both cities (**light orange** on maps), but much more widespread in Tallinn. Large parts of the Tallinn outer neighborhoods belong to this cluster, whereas in Helsinki only a small number does (but also in the outer parts of the city though).

This cluster is more intensive in frequent access to food places and stores to buy food.

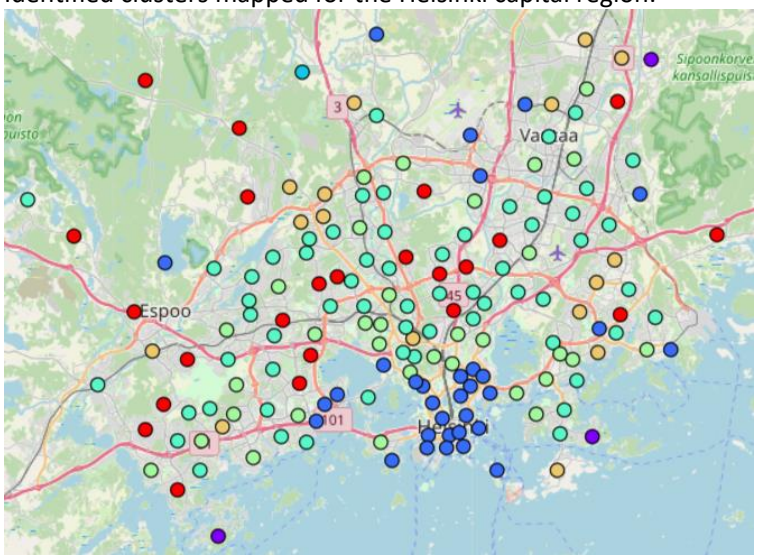
Labels	Venue Category	Freq
6	Grocery Store	0.066632
6	Supermarket	0.059283
6	Bus Station	0.040550
6	Bus Stop	0.038769
6	Fast Food Restaurant	0.038513
6	Pizza Place	0.032923
6	Café	0.031210
6	Gym	0.027997
6	Convenience Store	0.027521
6	Shopping Mall	0.027121

Cluster 7 is a middle sized cluster appearing only in Tallinn (**dark orange** on the map) and there mostly on the most southern side.

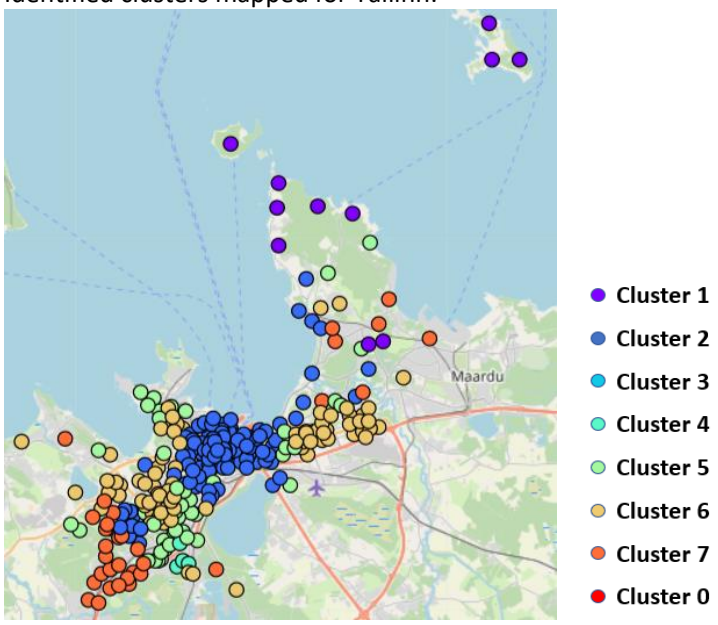
It this cluster, access to public transportation is also prominent, but the rest of the venue categories a show different profile ratio.

Labels	Venue Category	Freq
7	Bus Station	0.251918
7	Trail	0.058700
7	Grocery Store	0.041574
7	Train Station	0.035991
7	Park	0.033008
7	Bus Line	0.025504
7	Cable Car	0.020297
7	Restaurant	0.019938
7	Skate Park	0.019459
7	Café	0.015480

Identified clusters mapped for the Helsinki capital region:



Identified clusters mapped for Tallinn:



5. Discussion

The goal of this analysis has been to understanding the similarities of neighbourhoods between the Helsinki capital area and Tallinn in order to help people who consider moving from one city to the other.

For this end, the results of the k-Means clustering algorhitms have returned useful points to consider:

- Although the structural composition of the two cities is different, there are clear overlapping clusters:
    - The city centres and central locations of a few other city parts (cluster 2)
    - Neighborhoods with seashore and/or close to natural and historical sites (cluster 1, small)
- Moving futher outside from the city centres towards the suburbs, it becomes visible that more distant neighborhoods start getting a more and more diverging character in Helsinki and Tallinn:
    - Often encompassing the city centres, Cluster 5 appears and is significant in size in both cities.
    - In Helsinki, Cluster 4 seems to be a continuation to Cluster 5, representing an increased profile weight of public transportation and a decreased profile weight of shopping and dining venues. In Tallinn, Cluster 4 appears too, but it is very small and is located on the southern outskirts.
    - Similarly to Cluster 4 described above, Cluster 6 seems to be a similar position, only from a mirrored point of view. This cluster is very large in Tallinn, but rather small in Helsinki.
    - One step further to the external areas from both Cluster 4 and Cluster 6, Cluster 0 is unique for Helsinki and Cluster 7 is unique for Tallinn. There are clear differences in the profile category weights, but both clusters strongly pronounce public transportation. In Tallinn, Cluster 7 is mostly concentrated in the south-western corner. In the Helsinki capital region, Cluster 0 also tends to be in more external locations, but because of the city is very forest-intensive, this cluster also appears in locations closer to the centre in between other cluster segments.

Reviewing the administrative differences between the two countries was left out of scope. Similarly, diving deeper into the definition of Foursquare venue categories and potentially regrouping some of them was not considered for this project. However, it was still observed (and it is also visible in this report above) that there are some differences between the prevalence of some categories between the two cities. For example:

- There is a general ”Restaurant” category, strongly used in Helsinki but much less in Tallinn. On the other hand, there are quite many restaurant categories that come with type specification, such as ”Eastern European...” (unique for Tallinn) and ”Himalayan...” (unique for Helsinki). These are also restaurants, so it may be possible to group them together. Alternatively, it may be possible to identify the type of the restaurants from the general ”Restaurant” category. Both options may have advantages for the target group. Some people have strong preference for the type of restaurant, while others may not.
- The key word ”bus” appears in multiple categories, such as ”Bus stop”, ”Bus station” and ”Bus line”. Some of these categories even appear in the same clusters. Although these are likely to have different physical properties, their purpose still connects to the importance of public transportation. It may thus be possible to group them, given assumptions about the target group is taken and accepted.

The above points may be relevant if someone would venture to continue this clustering work in the future. At this point, these stay only as observations to extends the outcome of the neighborhood clustering assignment.

6. Conclusion

In this study, I analyzed the neighborhoods of the Helsinki capital region and Tallinn from the perspective of their top venue category profiles. I identified 8 clusters in total (practically speaking 7 clusters – as one was a remote cluster with a construction site). In this report, the clusters were described and shown on the maps of both cities. The Results and Discussion parts contain many points that may help those people who consider moving from one city to another and seek a neighborhood with a similar (or different) profile.

Final comment: The assignment also contained a Jupyter Notebook (written in Python 3.6) as assignment deliverable. The scope of this report is not to discuss the Python code itself, but to focus on the assingment from the analysis value point of view.