Clustering of Danish cities according to their similarity

Applied Data Science Capstone project

Alexander Kiilerich

January 16, 2020

Introduction

Background

Denmark is a small country in Northern Europe with a population of 5.6 million people.
Despite its moderate area of only 43,933km^2, the country is divided into a number of smaller regions, each containing several larger and smaller cities with very different characteristics.
At the same time, the danish job market is characterized by a high mobility and especially the large fraction of highly educated people change their workplace many times throughout their career.
This often necessitates the moving of families to different regions. It is beneficial to ease such transitions by seeking to move to a city or area in the new region which resembles ones previous place of living. However, this might not a priori be so easy since one might not have any experience from that particular part of the country. Therefore, it is desirable to apply data to compare cities across different of the country's regions in order identify which have similar characteristics; e.g. venues in a close vicinity.

Problem

Data on the geographical coordinates of each city above a certain size can be used to obtain information on the dominating venues in a given radius around that city from Foursquare. The project aims to use this combined data to cluster the danish cities according to their similarity in venue offerings. The main goal is to produce a map which can be used to quickly find cities in a given area of interest which are similar to towns or cities with which one has more experience. If possible, the individual clusters should be analyzed in order to label them according to their overall type (e.g., cultural cities, outdoor cities, commuter cities, etc.).

Stakeholders

First of all, the results would be interesting for people who are for whatever reasons set to move to a different part of the country. It might also be interesting for employers to offer this kind of tool as a help to new employees in order attract strong candidates from afar.
A completely different group of stakeholders could be businesses, planning to expand their brand with new stores or restaurants around the country. Here the ability to identify cities, which are similar to ones where they have branches that already do well, will be a good first indicator of where to open a new branch.

Data

Data sources

Data on the latitudes and longitudes of the 300 largest danish cities is available at http://www.tageo.com/index-e-da-cities-DK.htm. For instance the first three cities are listed as follows:

Rank	City	Population (2000)	Latitude (DD)	Longitude (DD)
1	**Kobenhavn**	1089700	55.680	12.570
2	**Arhus**	224400	56.160	10.210
3	**Odense**	145600	55.400	10.380

Including only the 300 largest cities is adequate for the purpose at hand as the smallest of these has only 1900 inhabitants. The first step will be to use pandas to scrape the location data from there. The site also includes data on the populations of each city. This data will be scraped as well. It might not be desirable to use in the clustering but will certainly be interesting when it comes to analyzing the clusters. The city coordinates will then be used to extract data on the venues in a given radius (maybe 10-30km) around each location via the Foursquare API.

A map of Denmark with markers indicating the cluster labels of each city will be generated by the Folium package in Python. In order to compare different regions of Denmark, a .json file containing the region boundaries is obtained from https://raw.githubusercontent.com/Neogeografen/dagi/master/geojson/regioner.geojson.

In an extension of the study it would be interesting to apply other datasets in addition the venue data in the clustering and description of the clusters. Of particular interest is, for instance, data on occupation, educational background, mean income and sports activities of the inhabitants in each city. Such datasets do not seem to be readily available online though, so the present analysis will focus on the venue data alone.

Data cleaning

The coordinate data from *tageo.com* was distributed over five pages, each containing data on 60 cities. Each of these were scraped and the data combined into a single dataframe. The first five entries of this frame are

	City	Population (2000)	Latitude (DD)	Longitude (DD)
0	Kobenhavn	1089700	55.68	12.57
1	Arhus	224400	56.16	10.21
2	Odense	145600	55.40	10.38
3	Aalborg	121500	57.03	9.93
4	Esbjerg	72500	55.47	8.45

The coordinate data was cleaned in two steps

1. At least one of the cities had NaN in the Latitude and Longitude columns. These entries were dropped from the data set.
2. To check the data for any clear errors, each city was marked on a world map using Folium. As seen in Fig. 1, the city Lemvig was listed with the wrong Latitude, placing it in Nigeria. We thus replaced the Latitude with the true value 56.5443 by hand.

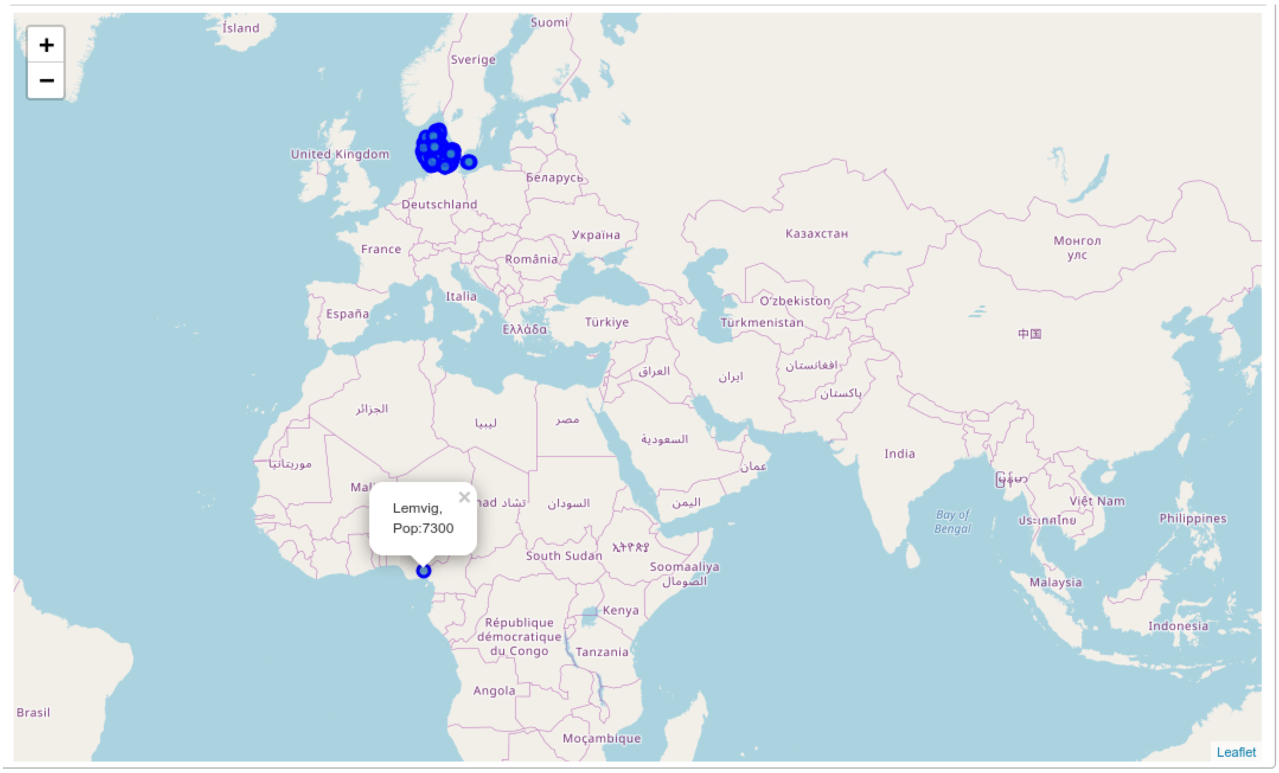


Fig. 1: The scraped cities shown on a world map.

Data preparation

For clustering

To prepare the venue data for clustering and further analysis, one hot encoding was performed on the dataset in order to turn categorical variables into numerical ones. The venues were then grouped for each city such that the dataset consisted of one row for each city with columns for each venue category. In order to facilitate an unbiased clustering, the final preparation step consisted in renormalizing the number of occurrences of each category to be represented as frequencies.

For region assignment

The dataframe with the city coordinates and venue data was turned into a geodataframe such that upon joining it with the geodataframe containing the region boundaries, each city would be assigned the correct Danish region.

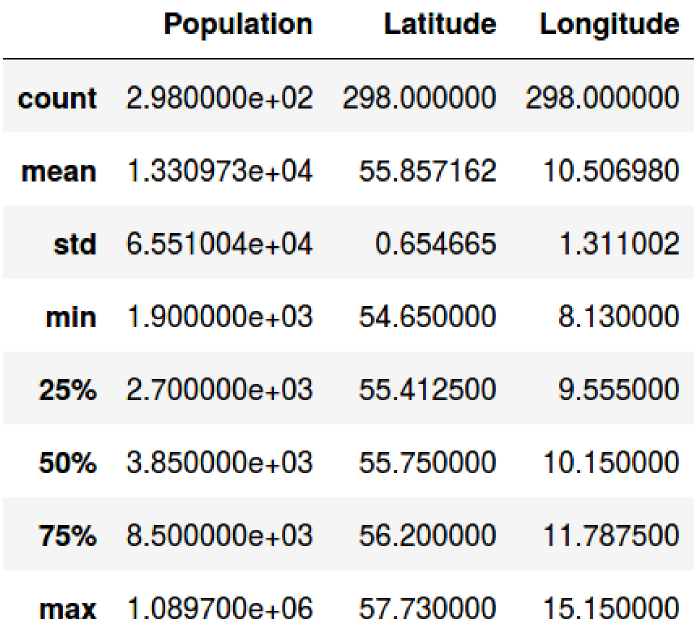
The final product was then a combined dataframe containing for each city:

- Population
- Coordinates
- Region
- Frequencies of most common venue categories
- The cluster label

Methodology

Exploratory data analysis

To acquire an initial insight regarding the demographic of the Danish cities, we first study the summary statistics of our population data. As seen in Table 1, the 298 largest cities have on average 13310 inhabitants, ranging between 1900 for the smallest Engesvang and 1089700 for the capital Copenhagen.



	Population	Latitude	Longitude
count	2.980000e+02	298.000000	298.000000
mean	1.330973e+04	55.857162	10.506980
std	6.551004e+04	0.654665	1.311002
min	1.900000e+03	54.650000	8.130000
25%	2.700000e+03	55.412500	9.555000
50%	3.850000e+03	55.750000	10.150000
75%	8.500000e+03	56.200000	11.787500
max	1.089700e+06	57.730000	15.150000

Table 1: Summary statistics for Danish cities

The regions with cities assigned different colors are shown on the map in Fig. 2

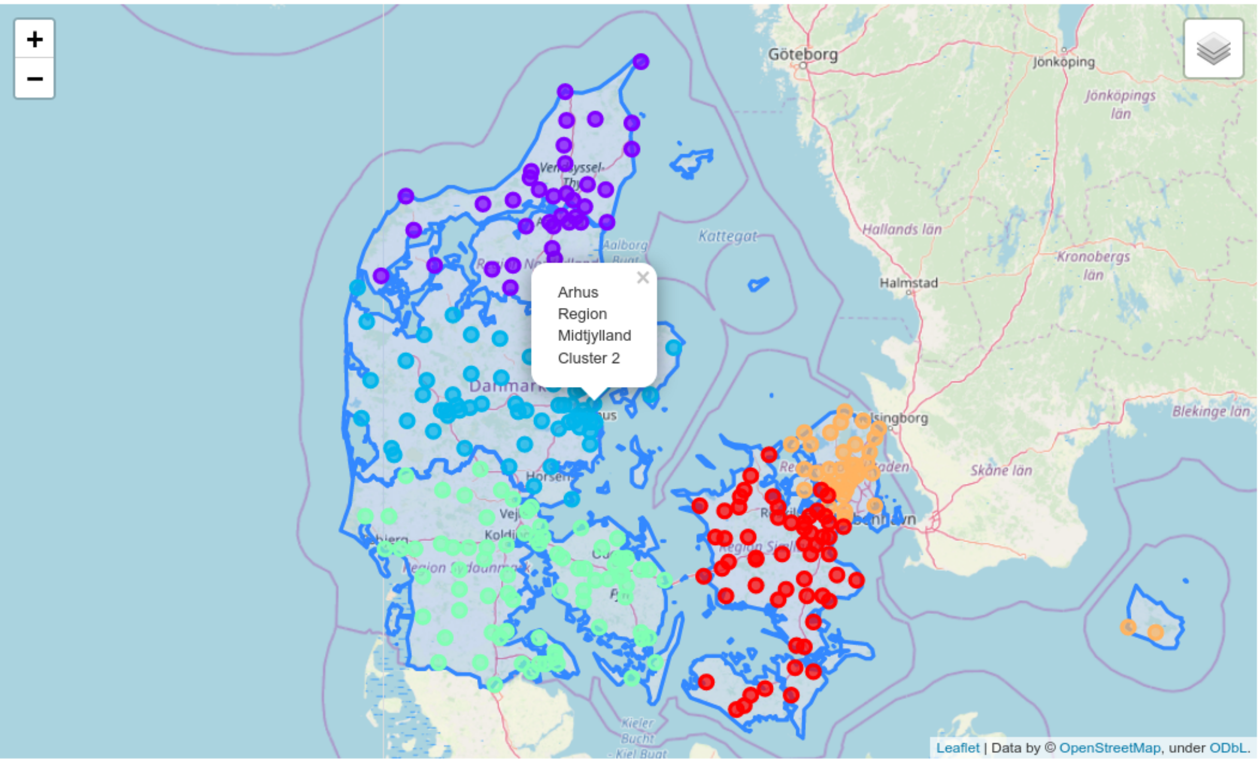


Fig. 2: Map of Denmark with cities sorted by color according to their regions. Orange: Region Hovedstaden. Light blue: Region Midtjylland. Purple: Region Nordjylland. Red: Region Sjælland. Green: Region Syddanmark.

We can compare the demographics across the regions by studying the barplot in Fig. 3. It is seen that the regions are similar in this respect. We note, however, that Region Hovedstaden (The Capitol Region) consists of more populated cities with the median population higher than the other regions.

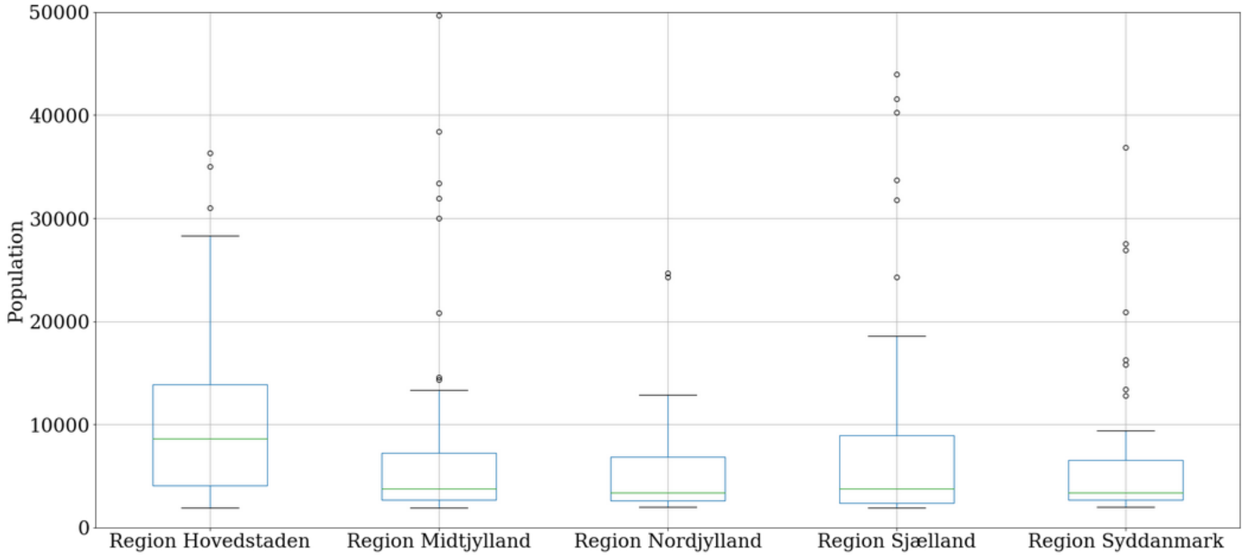


Fig. 3: Boxplot illustrating the population distribution across the five Danish regions. The population axis is truncated at 50000, since the major cities represent huge outliers which would otherwise hide the main parts of the plot.

Clustering according to venue data

We decided to included venues for each city in a radius of 10km around its centre. This number can be adjusted according to the specific wishes of the end-user and should be included as a parameter if this kind of analysis is developed into a distributed tool.

Upon some exploration, we found that reasonable converges in the clustering results was obtained by including only the 30 top venues for each city, and we decided to cluster the cities into 5 different classes.

The clustering of the cities according to their most frequent venues was performed by the K-means algorithm.

Categorization of clusters

After the clustering, the clusters were characterized by studying the five most common venues in each city belonging to a given cluster. When combining this with population data from each cluster and the geographical locations of each city, it was possible to produce a small description of each cluster.

Results

The results of the clustering are depicted on Fig. 4. The map is interactive with pop-up labels giving the region and cluster of each city. This map as a tool is the main product of this project.

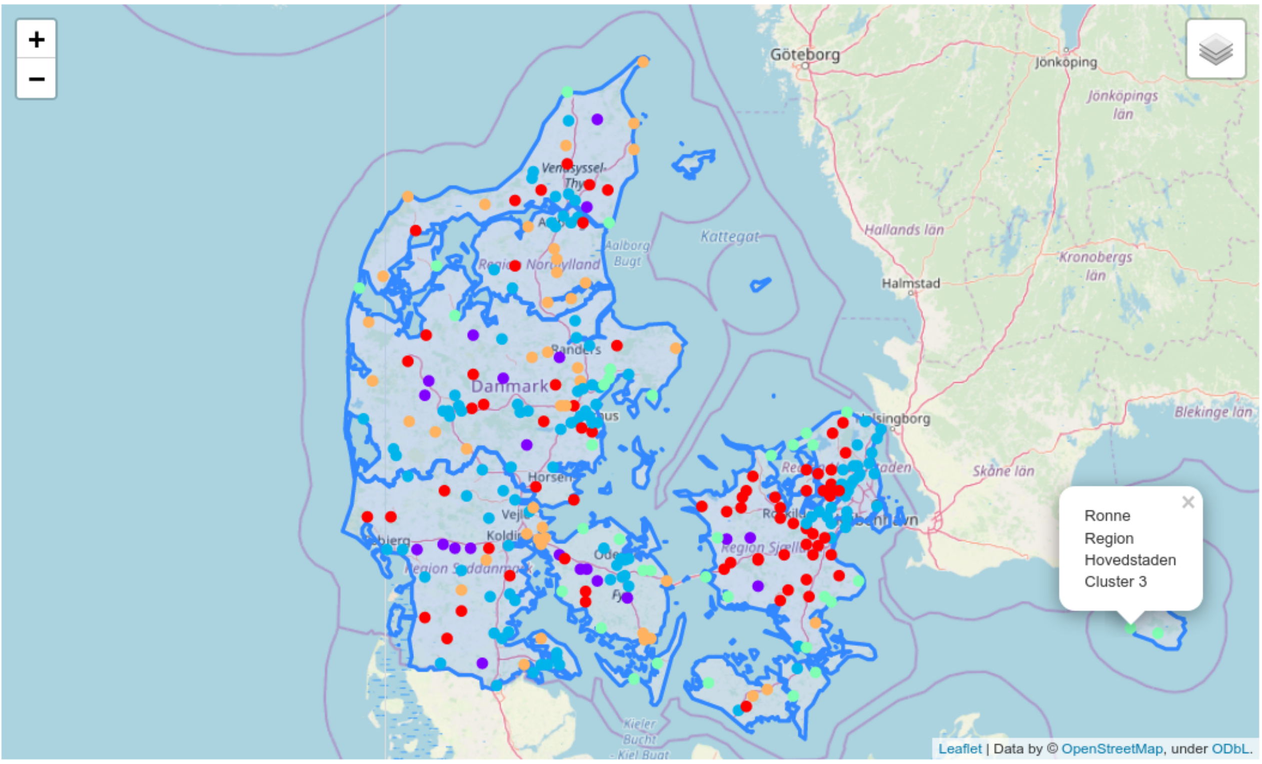


Fig 4: Map of Denmark with cities sorted by color according to their clustered class. Red: Cluster 0. Purple: Cluster 1. Blue: Cluster 2. Green: Cluster 3. Orange: Cluster 4.

In Fig. 5, we show the population distribution of each cluster.

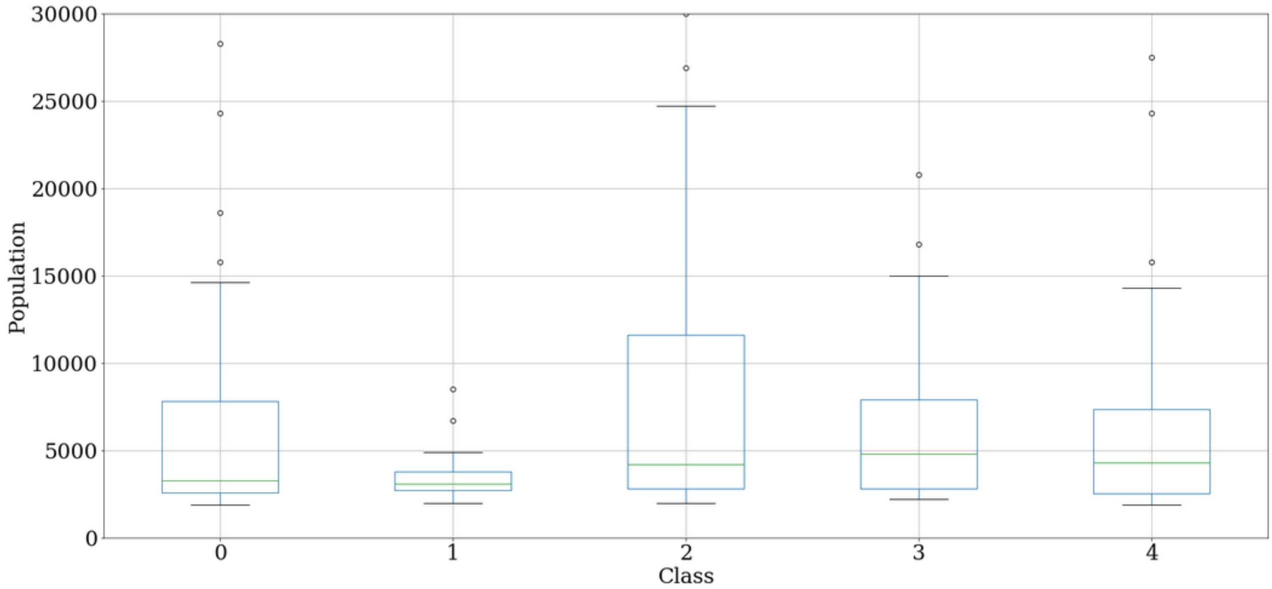


Fig. 5: Boxplot illustrating the population distribution across the five differerent clusters. The population axis is truncated at 50000, since the major cities represent huge outliers which would otherwise hide the main parts of the plot.

Discussion

Cluster map

We will first comment a bit on some of the immediate observations from the map in Fig. 4 with the clustered cities. We see that each region contains cities from a number of different clusters. The Capitol region (Region Hovedstaden) is dominated by cities in Cluster 2 which is also the cluster that most of the major cities (Aarhus, Odense, Aalborg, Esbjerg Herning, Randers, Silkeborg) are assigned to. The region around the Capitol area (Region Sjælland) is predominately assigned to Cluster 0.

Population distribution in clusters

Cluster 1 has a very narrow distribution with a low median value, and in line with the observation above from the map, Cluster 2 contains many large cities. Clusters 0, 3 and 4 are similar when it comes to the population distribution.

City categories

After the clustering, we may study the cities assigned to each cluster (an example for for Cluster 3 is shown in Table 2) in order to describe the general characteristics of each cluster and give them descriptive titles. Such an analysis leads to the following description of the five clusters:

Cluster 0 (red markers in Fig. 4):

> 72 towns with a mean population of 7272. The by far dominating venues are grocery stores. This cluster represents smaller towns from where people mainly commute to a larger city whenever they want to eat out of attend cultural or other activities. The towns themselves are mainly used for living and shopping of everyday goods and food. From Fig. 4, we note that these towns are predominately located in rings around larger cities (Cluster 2, blue markers), which offer a rich variety of venue offerings in close driving distance.

Cluster 1 (purple markers in Fig. 4):

> 21 towns with a mean population of 3542. These are small towns which all have train stations. The cities in this cluster are characterized as commuter towns where people in general need to commute whenever they want to go to work, eat out or enjoy any cultural activities.

Cluster 2 (blue markers in Fig. 4):

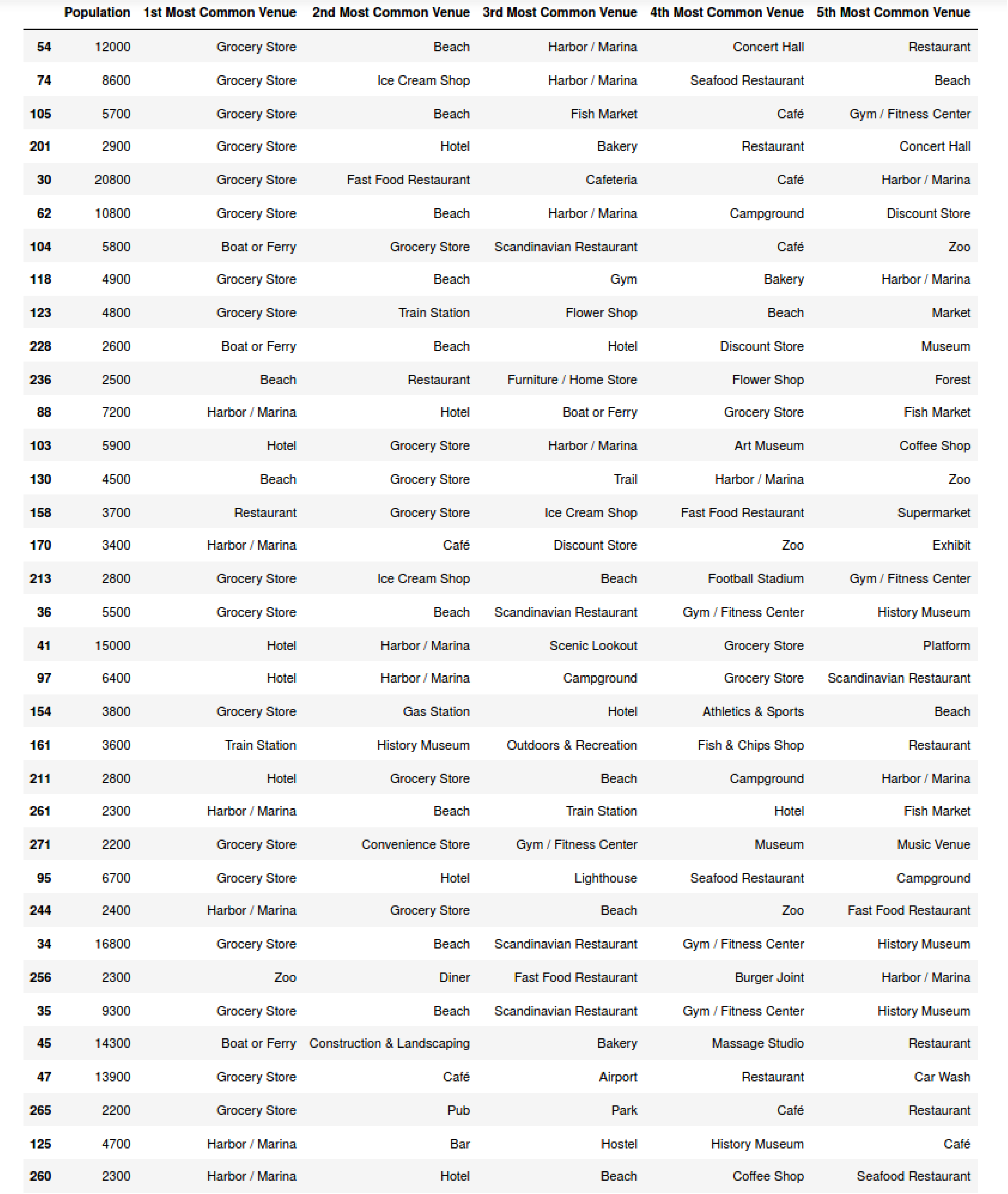
> 110 cities with a mean population of 24338. These are larger cities with a wide variety of offering for activities such as fitness centers, parks, golf courses, pools as well as many different cafés and restaurants.

Cluster 3 (green markers in Fig. 4):

> 35 towns with a mean population of 6440. These are smaller towns, very similar to those in Cluster 0 with grocery stores as the dominating venues. The difference is that the cities in Cluster 3 all have harbors or marinas, and many of them feature beaches. Looking af Fig. 4 we note that these cities are all located at the seaside.

Cluster 4 (orange markers in Fig. 4):

43 towns with a mean population of 6923 . These are smaller towns with some variety amongst them. One common thing is that they almost all have hotels as their most common venue and many of them feature gas stations. At the same time we note that there are many attractions (aquariums, historic sites, zoos, theme parks, etc.) among their common venues. From this, we infer that this cluster represents (mainly) cities where tourism and visitors play a large role.



	Population	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
54	12000	Grocery Store	Beach	Harbor / Marina	Concert Hall	Restaurant
74	8600	Grocery Store	Ice Cream Shop	Harbor / Marina	Seafood Restaurant	Beach
105	5700	Grocery Store	Beach	Fish Market	Café	Gym / Fitness Center
201	2900	Grocery Store	Hotel	Bakery	Restaurant	Concert Hall
30	20800	Grocery Store	Fast Food Restaurant	Cafeteria	Café	Harbor / Marina
62	10800	Grocery Store	Beach	Harbor / Marina	Campground	Discount Store
104	5800	Boat or Ferry	Grocery Store	Scandinavian Restaurant	Café	Zoo
118	4900	Grocery Store	Beach	Gym	Bakery	Harbor / Marina
123	4800	Grocery Store	Train Station	Flower Shop	Beach	Market
228	2600	Boat or Ferry	Beach	Hotel	Discount Store	Museum
236	2500	Beach	Restaurant	Furniture / Home Store	Flower Shop	Forest
88	7200	Harbor / Marina	Hotel	Boat or Ferry	Grocery Store	Fish Market
103	5900	Hotel	Grocery Store	Harbor / Marina	Art Museum	Coffee Shop
130	4500	Beach	Grocery Store	Trail	Harbor / Marina	Zoo
158	3700	Restaurant	Grocery Store	Ice Cream Shop	Fast Food Restaurant	Supermarket
170	3400	Harbor / Marina	Café	Discount Store	Zoo	Exhibit
213	2800	Grocery Store	Ice Cream Shop	Beach	Football Stadium	Gym / Fitness Center
36	5500	Grocery Store	Beach	Scandinavian Restaurant	Gym / Fitness Center	History Museum
41	15000	Hotel	Harbor / Marina	Scenic Lookout	Grocery Store	Platform
97	6400	Hotel	Harbor / Marina	Campground	Grocery Store	Scandinavian Restaurant
154	3800	Grocery Store	Gas Station	Hotel	Athletics & Sports	Beach
161	3600	Train Station	History Museum	Outdoors & Recreation	Fish & Chips Shop	Restaurant
211	2800	Hotel	Grocery Store	Beach	Campground	Harbor / Marina
261	2300	Harbor / Marina	Beach	Train Station	Hotel	Fish Market
271	2200	Grocery Store	Convenience Store	Gym / Fitness Center	Museum	Music Venue
95	6700	Grocery Store	Hotel	Lighthouse	Seafood Restaurant	Campground
244	2400	Harbor / Marina	Grocery Store	Beach	Zoo	Fast Food Restaurant
34	16800	Grocery Store	Beach	Scandinavian Restaurant	Gym / Fitness Center	History Museum
256	2300	Zoo	Diner	Fast Food Restaurant	Burger Joint	Harbor / Marina
35	9300	Grocery Store	Beach	Scandinavian Restaurant	Gym / Fitness Center	History Museum
45	14300	Boat or Ferry	Construction & Landscaping	Bakery	Massage Studio	Restaurant
47	13900	Grocery Store	Café	Airport	Restaurant	Car Wash
265	2200	Grocery Store	Pub	Park	Café	Restaurant
125	4700	Harbor / Marina	Bar	Hostel	History Museum	Café
260	2300	Harbor / Marina	Hotel	Beach	Coffee Shop	Seafood Restaurant

Table 2: Populations and 5 most common venues for cities in cluster 3-

Outlook

In an extension of this study it would be interesting to apply other datasets in addition the venue data in the clustering and description of the clusters. Of particular interest is, for instance, data on occupation, educational background, mean income and sports activities of the inhabitants in each city. Including such features in the characterization of the cities would provide a more complete picture which would allow for a more sophisticated clustering and hence a better accuracy for the end-users.

If the tool turns out to be successful, it would be straightforward to extend it outside the borders of Denmark. Both by making similar tools in other countries but also in order to be able to compare regions between two different countries, in case the upcoming move is to a different country.

Conclusion

The prospect of moving to or opening a business branch in a new region of Denmark can be eased by first identifying cities in the new area which have similar characteristic to ones with which one already have experience. By using venue data from the Foursquare database in combination with location data on danish cities, this project has made the initial steps in developing a tool for this kind of task. The product is an interactive map of Denmark where the 298 largest cities (with populations of at least 1900) are marked and assigned categories according to their similarity in local venue offerings. We found five distinct classes of cities (towns) which were each giving a small description according to the associated cities. In combination with the interactive map this should provide a valuable tool for the stakeholders.

Finally, we note that this project holds a large potential for further development in different directions. It would be very interesting to include different data than just that of venues. At the same time, the scope and applications of the tool proposed here extends far beyond the borders of Denmark so it would be obvious to generalize the analysis to other countries.