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 jobmarket 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 generated by the Folium package in Python.

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 main analysis will focus on the venue data alone.