Comparing Neighborhoods of Darmstadt and Karlsruhe: A data science study

Final project of the
‘Applied Data Science Capstone’
course on coursera.org

- Business Problem and Data Sources -

Dr.-Ing. Steffen Salenbauch

01-02-2019

Contents

1. Introduction / Business Problem

Did you ever move from one city to another? A recent article of a big real estate portal reveals that about 8.4 mio people in Germany move every year [1]. This corresponds to about 10% of the overall population of the country. While 75% have private reasons for a move, 25 % move for professional reasons [1]. It is conceivable that people from the first group move since they e.g. decide to live together, separate, get children or the children become older and move out. Probably, most of the people of this group only move within the same city or region. In the second group, however, people often move to other cities which might be hundreds of kilometers away from their hometown. Typical examples are young professionals who just finished their studies and get their first job, or specifically qualified employees who accept a job offer in another company.

This situation might be exciting but could also lead to uncertainty and doubts if one is not yet familiar with the new town. Which districts are safe, where are the areas with the huge shopping centers and where are the residential areas that are preferred by certain groups, e.g. families? As explained in the first part of the online course *Applied Data Science Capstone* on coursera.org, data science can help to answer these questions and reduce the uncertainty. Based on data from a local search-and-discovery service such as *Foursquare*, one can identify the most frequent venue types in certain neighborhoods and this information gives indication about the atmosphere and flair of the respective areas. For example, if a particular neighborhood has several university buildings, student dormitories and copy shops, one gets a clear idea about the everyday life in such a district.

This study focuses on the comparison and analysis of neighborhoods in the two german cities Karlsruhe (KA) and Darmstadt (DA). Considering german standards, both are of medium size (Darmstadt: about 161,000 [2] and Karlsruhe: about 310,000 inhabitants [3]). The two of them are well-known for their technical universities and several major companies, such as the pharmaceutical company Merck KGaA in Darmstadt or the electric utilities company EnBW Energie Baden-Württemberg AG.

The specific choice of these two cities is related to the personal, future plans of the author. He has just finished his PhD at the TU Darmstadt and got a job at a company in Karlsruhe. Hence, he is currently on the hunt for a nice flat in Karlsruhe. Since he is familiar with Darmstadt, he has a clear idea about the flair of the different districts. Thus, the current data science project aims to create a hierarchical, agglomerative clustering model for the neighborhoods of DA and KA to find which neighborhoods are similar. The idea is to get information which supports the author’s flat hunt activities.

To sum up, this data science study is of interest for people who either move from on of the two cities to the other one, or to data scientists, who are interested in different strategies in data science clustering studies in general.

2. Data

The study is conducted using data that is scraped from different sources in the web. As a first step, a complete list of all boroughs, neighborhoods and the corresponding population densities of Darmstadt and Karlsruhe are scraped from the following Wikipedia pages:

- Darmstadt: https://de.wikipedia.org/wiki/Liste_der_Stadtteile_von_Darmstadt,
- Karlsruhe: https://de.wikipedia.org/wiki/Liste_der_Stadtteile_von_Karlsruhe.

Based on the neighborhood names, the Python library geoPy is then used to determine the latitude and longitude coordinates of each neighborhood. However, since not all of the neighborhoods can be recognized correctly by geoPy, some coordinates need to be added manually. This is done with the support of the websites

- https://www.gpskoordinaten.de,
- https://tools.wmflabs.org/geohack.

The information on the neighborhoods' coordinates is further applied to query all venues within a certain radius around the centers of the neighborhoods from the local search-and-discovery service *Foursquare* using the respective application programming interface on https://developer.foursquare.com. Together with the information on the population densities, the wrangled venue data represents the features for the hierarchical cluster model introduced in the this study.

References

[1] ImmobilienScout24: Eine Infografik zum liebsten Hobby der Deutschen – dem Umzug. URL: https://blog.immobilienscout24.de/infografik-der-deutschen-liebstes-hobby/. Checked on 28-01-2019.

[2] Entwicklung der Bevölkerung in Darmstadt, Darmstadt-Dieburg, Hessen und Deutschland. URL: https://www.darmstadt.de/fileadmin/PDF-Rubriken/K02-1.pdf. Checked on 28-01-2019.

[3] Aktuelle Karlsruher Kennzahlen. URL: https://www.karlsruhe.de/b4/stadtentwicklung/statistik/kennzahlen.de. Checked on 28-01-2019.