Comparing Neighborhoods of Darmstadt and Karlsruhe: A data science study

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

- Full Report -

Dr.-Ing. Steffen Salenbauch

01-02-2019

Contents

1. Introduction

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.

3. Methodology

In order to determine which neighborhoods of Darmstadt and Karlsruhe are similar to each other, information about the venues localized within a radius of 800 m around the center of each neighborhood is scraped from Foursquare. Together with the information on the respective population densities of the neighborhoods, a hierarchical cluster model is developed. Illustrating the results in terms of a dendrogram allows to determine which neighborhoods are similar to each other answering the major questions of this study.

After conducting the data scraping steps explained in Section 2, a relational data frame is given, in which the observations are composed of the specific venue's name, the city name, the borough, the neighborhood, the population density, the geographical coordinates of both the center of the neighborhoods and the venue itself, and the venue's category. To transfrom the text attributes of the venue names to numbers, the method of

one-hot encoding is applied based on the venue category to get binary information about the type of each venue (‘0’: current venue does not belong to category X, ‘1’: current venue belongs to category X). The venues are then grouped to determine the number of occurrence of venues of each category in a certain neighborhood. Instead of storing the absolute numbers, the respective relative ratio R, defined by

R=number of venues of category X in neighborhood Z / overall number of venues in neighborhood Z,

is evaluated and stored.

As the features that will be used to create the cluster model not only include the information on the venues, but also the population density – which is a useful information to describe the character of a neighborhood – another data transformation step is required. This is related to the fact that the values of the venues all scale between [0,1], but the population density is given in absolute quantities (unit: inhabitants/ha). This would overemphasize the population density data with respect to the venue data in the clustering algorithm [4] and this should be avoided. Thus, the population data is rescaled to [0,1] using the global minimum and maximum values of its absolute quantity.

After these preparation steps, a hierarchical/agglomerative cluster model is developed using Python’s SciPy library. This method is useful to answer the fundamental question about the similarity of certain neighborhoods in the two cities. The model is finally illustrated in terms of a dendrogram. Based on this visualization, several conclusions can be drawn about the reasonable number of clusters or the similarity of certain neighborhoods in the two cities under investigation.

4. Results

The results of the hierarchical cluster model are visualized in the dendrogram in Fig. 1. It reveals that the data can be clustered into 2 groups, since there is a long distance between distance values 1.2 to about 2.8. However, in this study, we put the cutoff line to the distance 1.0 and get 3 clusters, which might also be a fair choice.

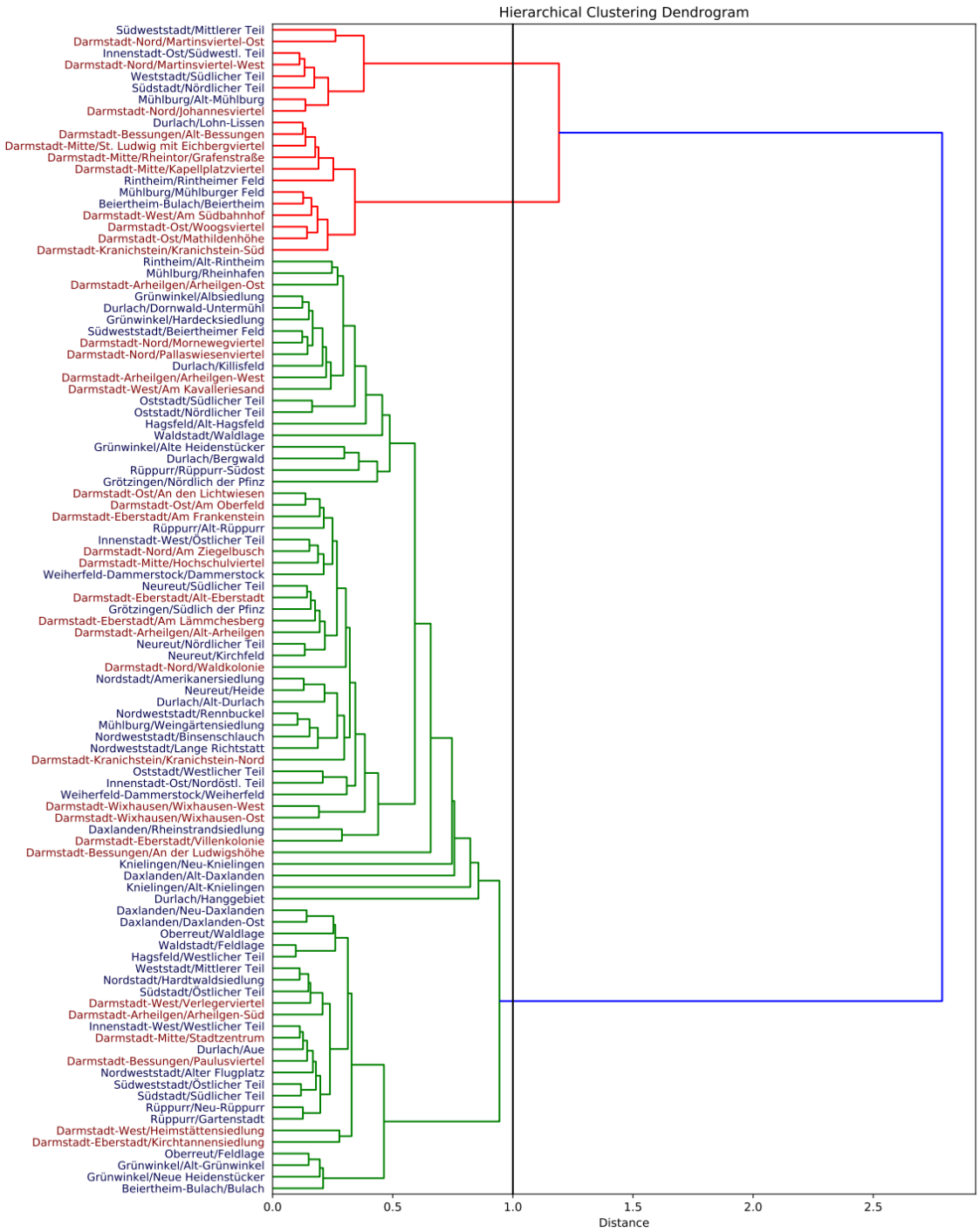

Figure 1: The dendrogram of the hierarchical cluster model. The black vertical line marks the chosen cutoff line, leading to 3 overall clusters. Brown labels on the y-axis mark neighborhoods from Darmstadt, blue mark the ones from Karlsruhe.

The dendrogram is helpful to determine which neighborhoods are similar to each other. For example, very similar neighborhoods have a rather small distance value (see x-axis in Fig. 1). Consequently, we can conclude that the most similar neighborhoods belonging to different cities are Martinsviertel-West (DA) and Innenstadt-Ost/Südwestl. Teil (KA).

Another major question of the study is to get to know which neighborhood in Karlsruhe is similar to Darmstadt-Nord/Pallaswiesenviertel, which is the current home neighborhood of the author. From the dendrogram, it can be noticed that the most compatible neighborhood in Karlsruhe is Südweststadt/Beiertheimer Feld.

In order to analyze the character of the neighborhoods of certain clusters, particular attributes are plotted in terms of scatter plots in Fig. 2. The attributes selected by the author are the scaled population density, the frequency of offices, hotel, bars, bakeries and playgrounds, as these quantities are helpful to evaluate the specific flair and atmosphere in the neighborhoods of certain clusters.

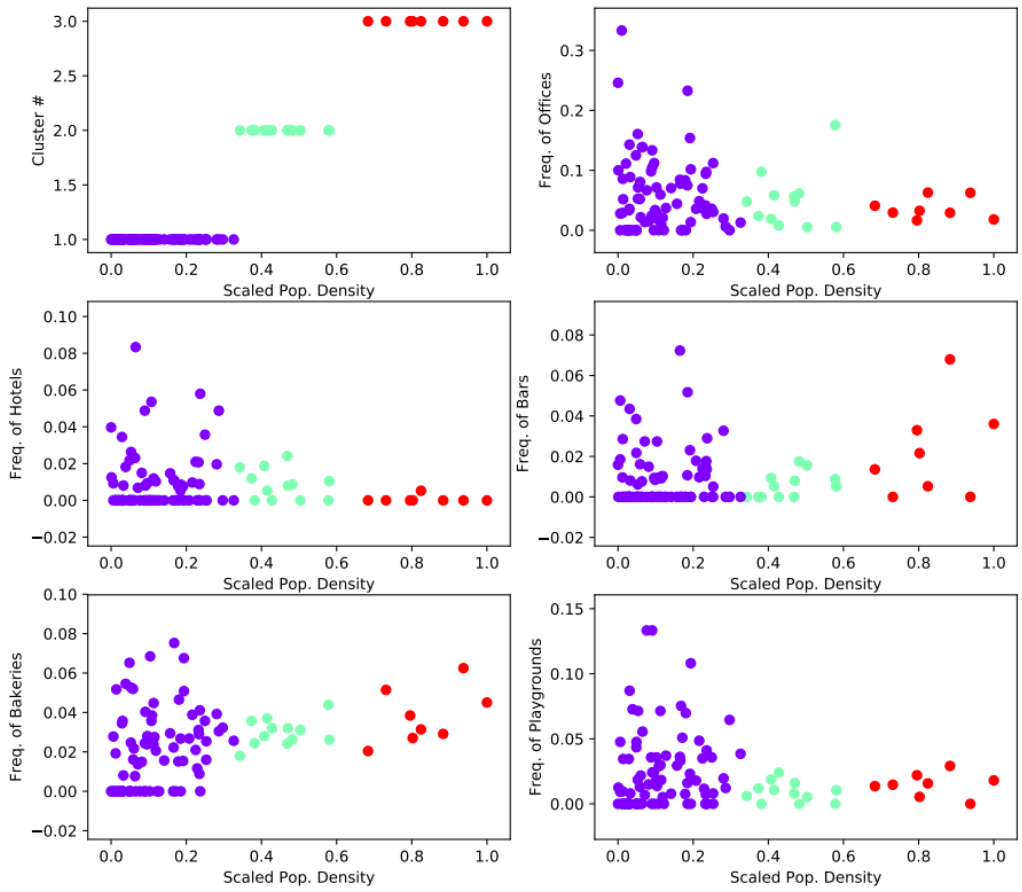

Figure 2: Scatter plots to visualize certain characteristics of the neighborhoods (represented by sample points). Clusters are highlighted by colors (cluster 1 in violet, cluster 2 in green and cluster 3 in red).

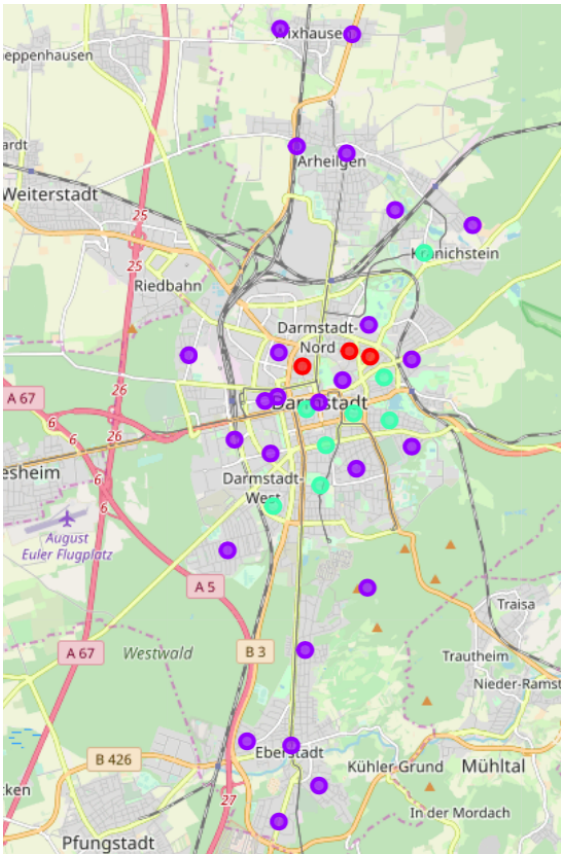

Figure 3: Map of Darmstadt with all neighborhoods as markers. Colors indicate clusters.

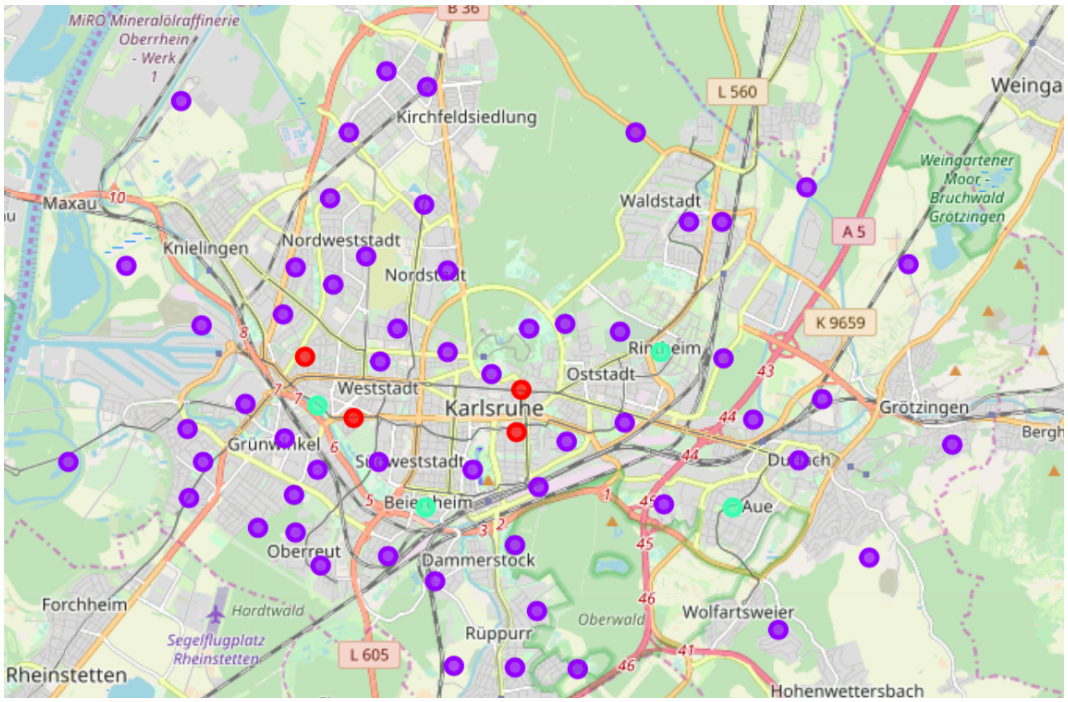

Figure 4: Map of Karlsruhe with all neighborhoods as markers. Colors indicate clusters.

Figures 3 and 4 visualize the locations of the various neighborhoods highlighting the respective clusters in different colors (keeping the same colors as in Fig. 2).

The findings described above are further discussed and analyzed in the following section.

5. Discussion

In this section, the results presented in Section 4 are analyzed to answer the fundamental questions of this study.

First of all, the dendrogram reveals that Martinsviertel-West (DA) and Innenstadt-Ost / Südwestl. Teil (KA) are the most similar neighborhoods in the two cities. This is further discussed based on the bar plots given in Fig. 5, which illustrate the frequency of the 10 most common venues in these neighborhoods.

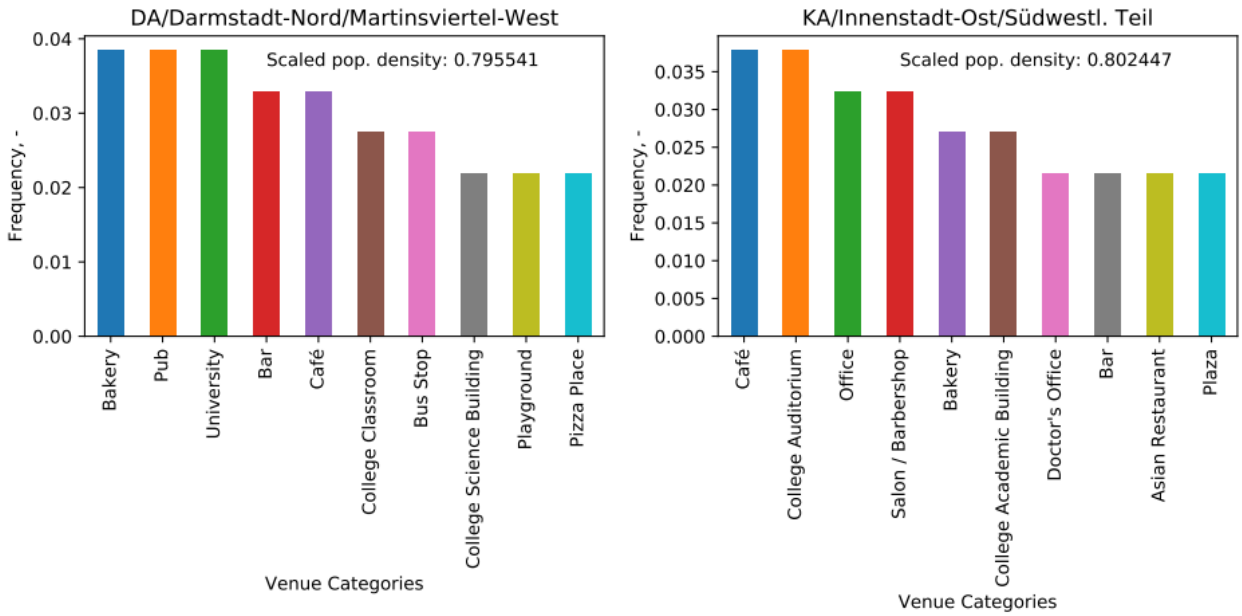

Figure 5: Frequency of the 10 most common venues of the two most similar neighborhoods of different cities Martinsviertel-West (DA) and Innenstadt-Ost / Südwestl. Teil (KA).

Both neighborhoods show a large population density. Together with the high frequency of bars/pubs, this highlights that both areas are residential areas, where people live and like to get together cozily in the evening. Furthermore, both areas have university/academic buildings nearby, suggesting that the neighborhoods are preferred among students. From

what the author knows about the Martinsviertel-West in Darmstadt, it is indeed an area that is very popular among young people and especially students. As mentioned in the introduction, the author is not yet familiar with Karlsruhe, however, a Google search revealed that Innenstadt-Ost is indeed popular among students, too.

Another major question is related to the personal situation of the author, who currently lives in Darmstadt/Pallaswiesenviertel, but will move to Karlsruhe soon. The dendrogram shows that Südweststadt / Beiertheimer Feld in Karlsruhe is the most similar neighborhood compared to Pallaswiesenviertel. This is again further analyzed having a look at the 10 most common venues in the two neighborhoods, see Fig. 6.

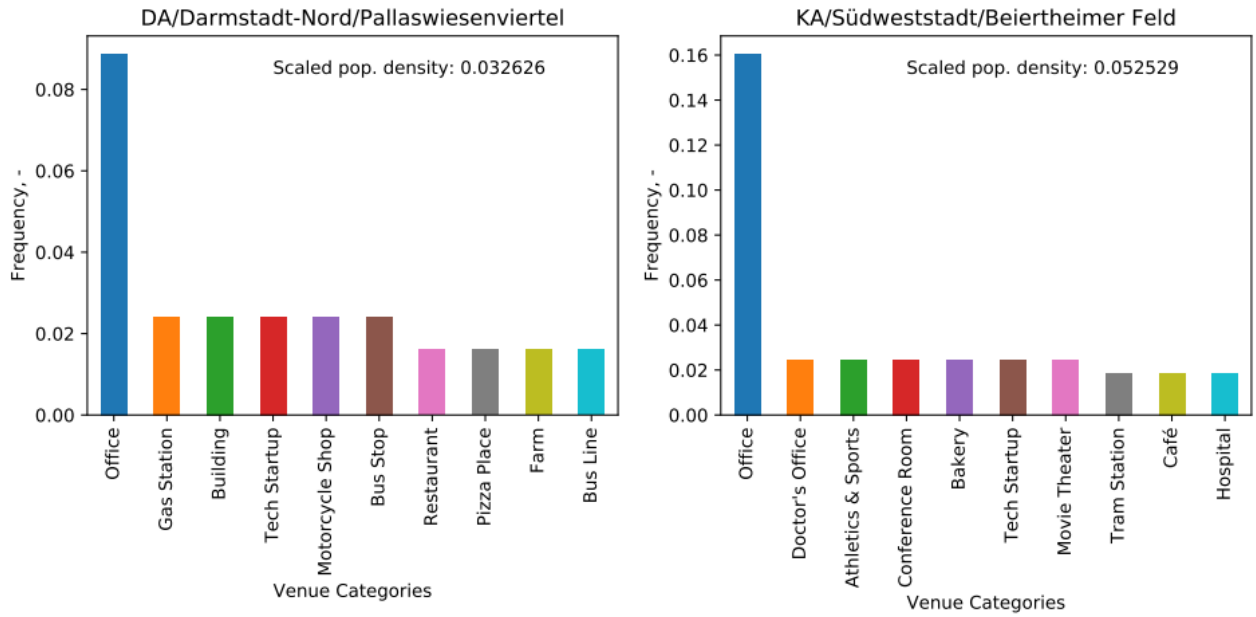

Figure 6: Frequency of the 10 most common venues of the neighborhoods Darmstadt/Pallaswiesenviertel and Karlsruhe/Südweststadt/Beiertheimer Feld.

Both neighborhoods reveal similar population densities with quantitative levels that are comparatively low. Together with the high frequency of offices in both areas, this clearly indicates that both neighborhoods are mainly commercial districts. This is further supported by the fact that bars and pubs are of minor importance here indicating that most people do not go out cozily in the evening to meet and drink a beer, but they are only around during the day to work.

Finally, it is interesting to discuss the characteristics of the 3 clusters selected visually from the dendrogram in Fig. 1. For this task, the scatter plots presented in Fig. 2 and the map plots in Figs. 3 and 4 were created. Based on this information, one could characterize the 3 clusters as follows:

Cluster 1:

Cluster 1 contains both residential and commercial shares. All neighborhoods have in common that the population density is rather low. However, the variance of the frequency of the attributes shown in Fig. 1 (offices, bars, playgrounds, …) of the neighborhood samples is quite large. For example, in this cluster, we have neighborhoods without any offices and ones with a large ratio of offices. We have to conclude that the current model cannot distinguish between commercial districts and residential districts with a small population density. The neighborhoods belonging to this category could either be commercial areas (large frequency of offices, hotels), residential areas on the edge of the town (where families live explaining the low population densities and the substantial frequency of playgrounds) or mixed areas. In order to better distinguish these groups, further work is required to improve the model in a future study.

Cluster 2:

Neighborhoods of cluster 2 are mainly residential areas. As most of the neighborhoods are rather on the edge of the cities (where space should be available), the small number of playgrounds suggests that these neighborhoods are preferred by people without children. The combination of the locations rather outside the city centers and the rather high population ratio further indicates many multi-storey buildings. Despite being residential areas, the aforementioned facts and the small frequency of bars leads to the assumption that these neighborhoods are popular among elder people.

Cluster 3:

Here, we have again residential areas for people without children and the presence of multi-storey buildings (see population density and playgrounds). The low frequency of offices and hotels, together with the high frequency of bars suggests that these neighborhoods might be preferred by students and rather young people. This is supported by the fact that the neighborhoods of cluster 3 are mainly downtown which is a popular location criterion among this group of people.

6. Conclusions

In this study, a hierarchical, agglomerative cluster model was developed to investigate the similarity of the neighborhoods of the two german cities Darmstadt and Karlsruhe. Visualizing the results in a dendrogram suggested that all neighborhoods can be described in terms of 3 distinctive cluster groups. Although these groups give insight into which areas might be preferred by young or elder people, it cannot distinguish between commercial districts and residential districts for families living on the edge of the cities. This aspect needs to be improved future updates of the model.

However, the main purpose of this study was to help the author and all other people who move from one of the cities to the other to get an impression which neighborhoods in the unfamiliar city are similar to the familiar one. This information can be determined from the dendrogram, which was created using the developed cluster model. It revealed that the most common neighborhoods in Karlsruhe and Darmstadt are the Martinsviertel-West (DA) and Innenstadt-Ost/Südwestl. Teil (KA). Both of them are areas that are popular among students. Furthermore, the it was found that Südweststadt/Beiertheimer Feld in Karlsruhe is the neighborhood which is most similar to Pallaswiesenviertel, which is the current home neighborhood of the author.

The entire Python code created in this study is freely available at https://github.com/SteSa1987/coursera-ibm-capstone.git.

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.

[4] Scardapane, Simone. Comment on researchgate.net. URL: https://www.researchgate.net/post/Does_normalization_of_data_always_improve_the_clustering_results. Checked on 31-01-2019.