Capstone Project IBM Data Science

by Patrick Geist

1. *A description of the problem and a discussion of the background*

My Capstone Project deals with the clustering of districts in Mannheim, Germany. I try to provide crucial insights to the individual districts concerning their similarity or dissimilarity. The core idea of this research is to use publicly available data and location data to cluster the districts of Mannheim. Upon the final clustering one can draw conclusion about the similarity and dissimilarity of districts which can provide crucial advice for several agents (new and existing restaurant owners, people with cultural businesses and those looking to expand in it, start-ups, the municipality of Mannheim and further). The conclusions to be drawn can be multifaceted. The analysis can serve as an ideal starting point for further customized more detailed analysis.

2. *A description of the data and how it will be used to solve the problem*

The analysis will make use of location data that are retrieved using the foursquare API plus publicly available demographic data that can be retrieved using the open data website of the municipality of Mannheim (https://www.mannheim.de/de/stadt-gestalten/daten-und-fakten/open-data). The demographic data that I will be using per district throughout the analysis will be:

- total population
- population growth (CAGR)
- age structure (6 different age groups)
- data from the last federal election to see party preferences across districts

Together with location data of venues in the respective district the data set can be described as a good starting point for clustering that includes multifaceted aspects of various districts in order to achieve a sophisticated clustering analysis.

3. *Methodology*

The main goal of this analysis is to establish a clustering of the districts in Mannheim based on the criteria explained in the data section. To achieve this, the respective data is pre-processed and standardized so that it can be used for the k-means clustering algorithm.

The clustering algorithm uses 5 centroids as this allows for enough variation across the districts to also point out some unique characteristics of some districts. The clustering is performed using the k-means clustering algorithm.

Furthermore, the clustering of the districts is conducted with respect to 3 different subsets. First, the clustering with the demographic data only. Second, with the venue data only. Third, with the two datasets combined. This procedure allows for a distinct analysis and offers the reader (or user) a holistic way to carry out his analysis of the different districts with respect to his preferences.

4. Results

From the clustering analysis one can draw the following conclusion: It seems that the city centre and its surrounding districts offer unique location with respects to venue and demographic data. However, it becomes obvious that when analysing the subsets that the population structure (with respect to demographics) is differently clustered distributed than that of the venue data. The complete dataset also leads to a different clustering.

It can be noted that the **clustering of venue data** and the **combined dataset** is more concentrated than in the other clustering. The clustering with venues is highly concentrated on the city-centre and the surrounding districts which confirms the intuitive notion that in the city centre there are more venues than in in any other district. It can also be observed that the clustering appears to move from inside to the outside. That is that as one further approach the city limits venues become scarcer.

For the **demographic clustering** it can be said that it is location dependent and there is a correlation between location and clustering which emphasizes the self-selection of different population structures in different districts.

5. Discussion

The results underline that there are crucial possibilities for new businesses in areas where a similar population structure suffers from a lower frequency of certain venues. In these areas (that are usually in the middle between the city-centre and the city limits) new businesses could thrive with an already confirmed POC (prove of concept) as similar venues exist in other districts with a similar population structure. This finding enables potential new businesses to select their location with respect to population structures that are proven to work for their business.

This finding should enable more efficient location selection of new businesses with respect to their target audience.

6. Conclusion

To sum up, this analysis establishes that there is crucial potential for new businesses to expand in other districts. It has been stated that this preposition can reduce the uncertainty involved in opening a new business (B2C) as the demand for the respective good has already been tested in another micro-market (alias district) that possesses the same population characteristics.