# Comparison of German Cities

Final project during Coursera course

„Applied Data Science Capstone“

Dr. Bernd Künnen

ARS Computer und Consulting GmbH

14.01.2019

# Content

Abstract ........ 2

Introduction ........ 3

The data ........ 3

Methodology ........ 3

Data preparation ........ 3

Machine Learing ........ 4

Results ........ 4

The Foursquare database ........ 4

The optimum number of clusters ........ 5

The clusters ........ 5

Discussion ........ 7

Appendix ........ 8

Cities that had to be excluded (num_venues<15): ........ 8

Calculating the optimum k ........ 8

Clusters ........ 9

Cluster 0 ........ 9

Cluster 1 ........ 9

Cluster 2 ........ 9

# Abstract

Many cities in East and North German are suffering from a loss of population while the cities in West and South are gaining residents. This may lead to the assumption that there're similarities between the cities in these two regions, respectively, while both groups are having significant differences .

To verify this assumption the 80 largest german cities have been compared based on data that's free available in the internet. The comparison was carried out looking at the occurences of public venues or points of interests, to see what cities have a similar „venue-fingerprint“ and if there're clearly shaped „clusters“ of cities in some parts of Germany.

# Introduction

Since the fall oft he wall, Germany experienced a unprecedented migration of people. Over the years many people moved from the eastgerman states to the western and southern parts of Germany[1]. In recent years and due to the fact that the most industrialised centers of Germany are in the South, another migration from North to South set in, though maybe not that distinct like the former one.

Most likely the reasons for migration from East to West is to be found in the labor market. Nevertheless, there could be additional reasons in the structure of the cities: A city that offers many venues such as cafès, bars, cinemas, theaters etc. is more likely to hold ist residents than a city that has a lower recreational value.

Therefore this project, performed during the Coursera course „Applied Data Science Capstone“, examines the „venue fingerprint“ of the german cities, based on public available data and driven by methods of Data Science.

# The data

For the project public available data was used. The basic list german cities was retrieved from a Wikipedia article about the largest german cities[2] which was then extended with the geological coordinates from Openstreetmap database.

The venue data fort he cities was retrieved from the Foursquare database, using their public API[3].

# Methodology

## Data preparation

In the first step a list of the largest cities in Germany including their geological coordinates was built. The Wikipedia article about the largest german cities provides[4] a list of 80 cities with more than 100.000 residents. This html table was scraped using Python and the BeautifulSoup library.

Subsequently for each city the geological coordinates for the city center was retrieved using the GeoPy library and thus refering to OpenStreetMap data. The city of Mönchengladbach had to be removed from the dataset because for unknown reason GeoPy delivered no coordinates.

[1] Demografische Situation in den ostdeutschen Ländern, https://www.beauftragter-neue-laender.de/BNL/Navigation/DE/Themen/Gleichwertige_Lebensverhaeltnisse_schaffen/Demografie/Demografische_Situation/demografische_situation.html
[2] Liste der Großstädte in Deutschland, https://de.wikipedia.org/wiki/Liste_der_Gro%C3%9Fst%C3%A4dte_in_Deutschland
[3]Foursquare API, https://support.foursquare.com/hc/en-us/articles/201219550-Foursquare-API-
[4] Liste der Großstädte in Deutschland, https://de.wikipedia.org/wiki/Liste_der_Gro%C3%9Fst%C3%A4dte_in_Deutschland

In the  second step a list of max. 100 venues was retrieved for each oft he 77 cities by using the Foursquare API[5], such as cafés, theaters, parks etc. The venues were retrieved for an area of 1000m around the city center, assuming that the greater part of the social life is taking place here and thus this area should be representative.

An overview was generated of how many venues had been retrieved per city. For that later for each city a fingerprint based on the 15 most occuring venue types will be built, all cities with less venue types had to be excluded from further processing. A list of these ten cities ist to be found in the appendix.

The data was then subsequently transformed into a dataset in which for each city the occurences of each of the 277 venue types was represented.

## Machine Learing

Based on this dataset, another reduced dataset was prepared which contained only the 15 most frequent venue types for each city, as a kind of „minimum fingerprint“.  The reduced dataset was then used for machine learning, where in a first step a k-means model was calculated for k=2 to k=15, trying to obtain the optimum for k. Based on the optimum k, the cities were assigned by k-means to different clusters.

# Results

## The Foursquare database

While preparing the data it was found that for most cities in Germany the Foursquare database doesn’t contain as much data as for cities in the US or Canada. As can be seen in the histogram below (Illustration 1), for about the half oft he cities that were investigated, less than 50 venues have been returned, whereas for large North-American cities a histogram would contain bars just to the right.

In detail for the center of both e.g. Bielefeld and Ludwigshafen am Rhein the API showed only 4 venues each, although these cities are on rank 18 and 46 in the base list, which is sorted descending by population. This indicates that the user base of Foursquare is not that big in Germany whereas in North America it seems to be widely used.

As it seems, data for many german cities ist provided only by few people. This may lead to a certain bias in the data. This point has to be taken in account when interpreting the results of the machine learning section.

[5] Foursquare API, https://support.foursquare.com/hc/en-us/articles/201219550-Foursquare-API-

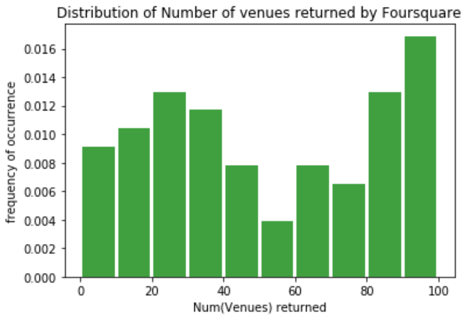

**Illustration 1: Histogram of number of venues returned by Foursquare for the 67 cities**

## The optimum number of clusters

In a first step oft he machine learning part, it was tried to determine the optimum number of clusters for the k-means algorythm which should be used subsequently. Therefore the k-means was performed for *k*=2 to *k* =14. The resulting costs are shown in teh appendix as well as in the plot „Illustration 2: Cost for different *k* on k-means clustering“. The plot doesn’t show any point with significant change in slope. The lowest number that shows a decent change in slope is *k* =3. Thus this number was chosen to use in the finally carried out clustering process.

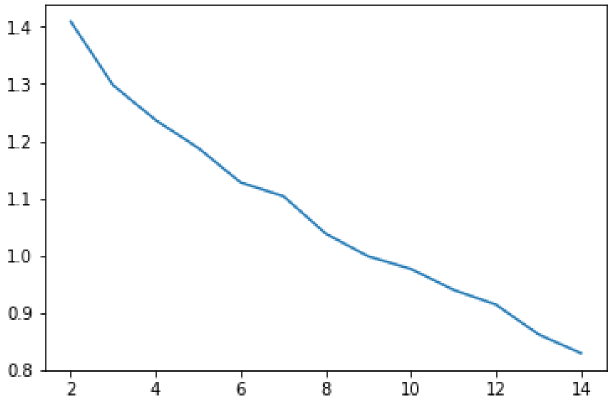

**Illustration 2: Cost for different *k* on k-means clustering**

## The clusters

Using the k-means method, the cities were assigned to five clusters. Surprisingly, more than 70% of the cities were assigned to one clusters, as can be seen in table below. The remaining cities were distributed almost evenly over the other two clusters.

| Cluster | Number of cities |
|---|---|
| 0 | 8 |
| 1 | 10 |
| 2 | 49 |

**Table 1: Clusters and the number of cities they contain**

An analysis on the geographical map of Germany (see Illustration 3: Map of Germany with superimposed a distribution oft he three clusters) shows an even distribution of the cities in the main clusters (red markers) whereas the cities of cluster 0 (yellow markers) are almost all in or near the so called „Ruhrgebiet“,  Germanies former industrial heart. The cities in cluster 1 (cyan markers) are spread over the south-west half of Germany but the distribution don’t seem to follow a certain pattern.

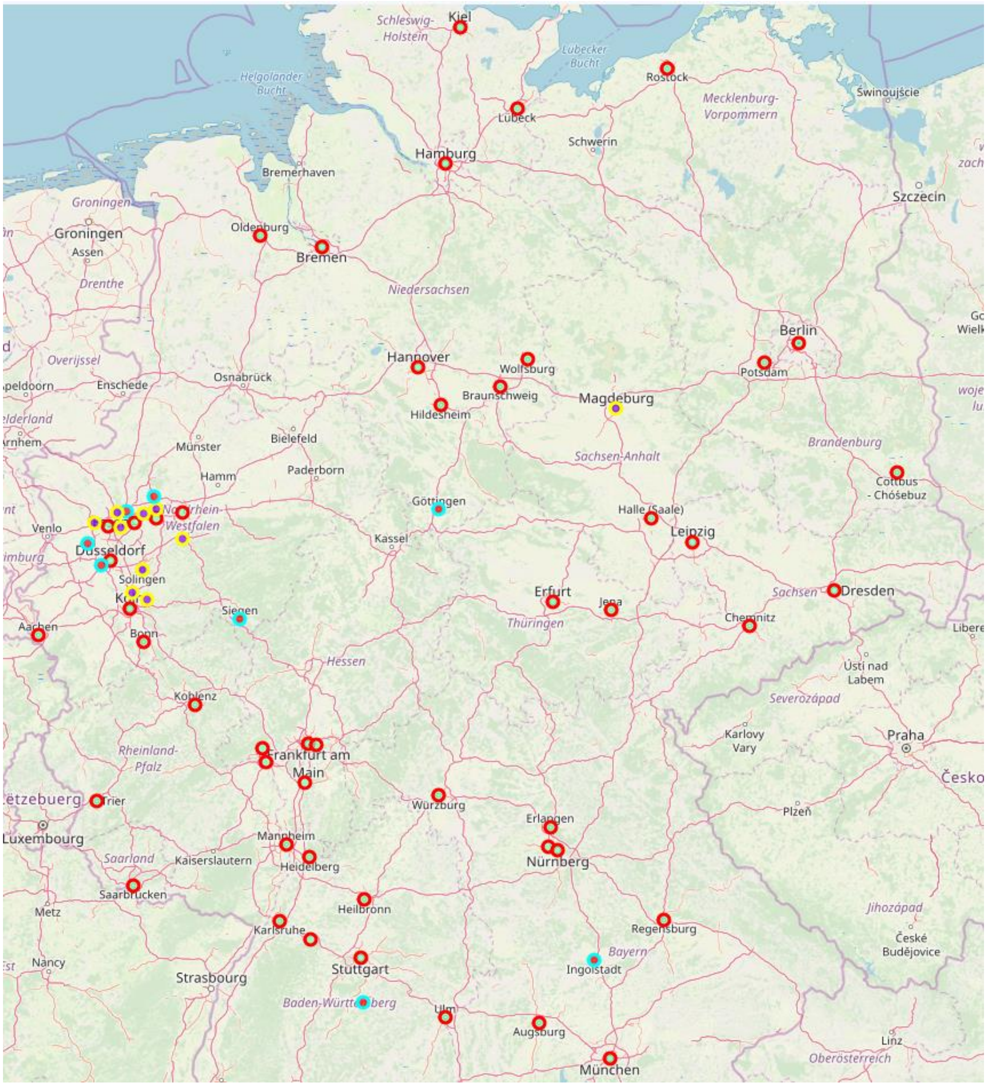

Illustration 3: Map of Germany with superimposed a distribution oft he three clusters

# Discussion

The result did not show a clear distinction of german cities according to some west/east distribution. On the contrary, most cities in Germany seem to have a similar „venue fingerprint“ and this group of cities is evenly distributed throughout Germany. So the assumption that there’s some obvious differences between cities in East and West Germany could not be proved.

A second, quite smaller cluster consists of ten if which nine are in the Ruhr region. This is the only cluster of cities that could be assigned to a dedicated area. The Ruhr area has a very distinct history, having been the indutrial heart of Germany for a long time and until the 1980s , based on coal and steel industry. Thus it might be that in these cities some special „Ruhr“ structure still is present, reflecting habits and social structures from former times.

The third cluster consists of eight cities that are tob e found over a wide area of west and south Germany. As there’re no obvious similarities to these cities, they seem to represent mainly a group of „outliers“. This is backed by the results of a test with distribution of the cities to six clusters where the additional three clusters contained just very few cities, with two of the clusters containing just one city. Most of these are part of the outlier-cluster in the three-cluster model.

Finally, some difficulties arose from the Foursquare database and ist API. First, the database seems to contain not enough data to make reliabe statements for a real world business project. For many cities the database contains less than 100 venue entries which is not representative. On the other hand, for big city the API delivers „only“ 100 venues, so at different times the given json data may contain different venue data which makes it hard to verify the final results. Furthermore this data is most likely gathered by just a german Foursquare users so that their personal habits may leed to some bias in the data.

Nevertheless, for a studying project and to answer the introductory question the data seems to fit. It was a good result to see that there doesn’t seem to be a significant difference between german cities. In a next project, the cluster with the Ruhr cities should further be investigated to understand what makes them a little bit special.

# Appendix

## Cities that had to be excluded (num_venues<15):

Bielefeld

Freiburg im Breisgau

Hamm

Kassel

Ludwigshafen am Rhein

Münster

Osnabrück

Paderborn

Remscheid

Wuppertal

## Calculating the optimum k

These are the results of calculating the costs for different numer of k-means clusters, trying to find the optimum *k*.

```
k: 2  cost: 1.40889691131
k: 3  cost: 1.29733019354
k: 4  cost: 1.23662397479
k: 5  cost: 1.18799402143
k: 6  cost: 1.12755921236
k: 7  cost: 1.10395357524
k: 8  cost: 1.03820725672
k: 9  cost: 0.998576633408
k: 10  cost: 0.976727851367
k: 11  cost: 0.940269545841
k: 12  cost: 0.914421655892
k: 13  cost: 0.862489751039
k: 14  cost: 0.829497892981
```

## Clusters

This section lists the three clusters and the cities they contain.

### Cluster 0

Bottrop
Göttingen
Ingolstadt
Krefeld
Neuss
Recklinghausen
Reutlingen
Siegen

### Cluster 1

Bergisch Gladbach
Gelsenkirchen
Hagen
Herne
Leverkusen
Magdeburg
Moers
Mülheim an der Ruhr
Oberhausen
Solingen

### Cluster 2

Aachen
Augsburg
Berlin
Bochum
Bonn
Braunschweig
Bremen
Chemnitz
Cottbus
Darmstadt
Dortmund
Dresden
Duisburg
Düsseldorf
Erfurt
Erlangen
Essen
Frankfurt am Main
Fürth
Halle (Saale)
Hamburg
Hannover
Heidelberg
Heilbronn
Hildesheim
Jena
Karlsruhe
Kiel
Koblenz
Köln
Leipzig
Lübeck
Mainz
Mannheim
München
Nürnberg
Offenbach am Main
Oldenburg (Oldb)
Pforzheim
Potsdam
Regensburg
Rostock
Saarbrücken
Stuttgart
Trier
Ulm
Wiesbaden
Wolfsburg
Würzburg