Report: Clustering cities and areas in cities based on their venue features

narges.poursaleh@gmail.com

February 2021

1 Introduction

1.1 Background

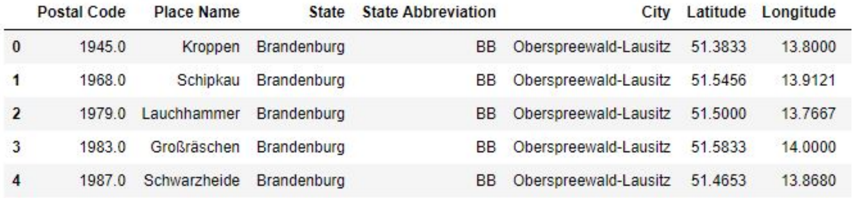
Neighborhood segmentation and clustering using geodata and foursquare API is an approach to apply many various business problems. For instance, discover a suitable place for opening a new shop or mall, building, or renting a house, hospital, school, etc. Notably, moving to a new home is one of the critical decisions that every family faces, particularly when they want to build a new house in a new city, that they may not know well ich very challenging and time-consuming. Having some neighborhood data as the sample, we can learn the pattern and find similar neighborhoods and recommend them to the person looking for a suitable place.

1.2 problem

Finding the best-fit neighborhood in the city is a well-known problem for people moving from one city to another. A friendly area differs from person to person since different persons or families have other criteria for choosing a neighborhood to live in. among different criteria for selecting a community to live in is the venues in an area. Some people prefer to live in a city center where there are lots of cafes and restaurants. Others may like some quiet places beside parks or a combination of facilities or venue categories near them. Finding an area that has all the criteria may be challenging. Sometimes people may like a neighborhood with the combination of its characteristics, but they wouldn't want to name all those properties. For example, a person would like to move to B city. She looks for a neighborhood like her current home in city A. without describing her criteria for living in an area. We aim to cluster two cities' areas to find a similar place in cities A and B. The result could be considered as a recommended resource for recommending similar locations in different cities. Further, we cluster ten other cities to categorize the cities into three different groups.

2 Data sources

We choose the data of cities in Germany to work with. We collected the towns' coordinates and postal codes from Aggadata to download all cities' geodata for free. As we couldn't find the neighborhoods' names, we decided to consider the postal codes to identify our neighborhoods. The data sources consist of postal codes and longitude and Latitude, city name, and the city's state. We collected the venues' names and categories together with each postal code's venus coordinates using the Foursquare API. Coordinate of cities center are obtained using Google Maps. In the figure 1 we have shown some example records in our dataset.



	Postal Code	Place Name	State	State Abbreviation	City	Latitude	Longitude
0	1945.0	Kroppen	Brandenburg	BB	Oberspreewald-Lausitz	51.3833	13.8000
1	1968.0	Schipkau	Brandenburg	BB	Oberspreewald-Lausitz	51.5456	13.9121
2	1979.0	Lauchhammer	Brandenburg	BB	Oberspreewald-Lausitz	51.5000	13.7667
3	1983.0	Großräschen	Brandenburg	BB	Oberspreewald-Lausitz	51.5833	14.0000
4	1987.0	Schwarzheide	Brandenburg	BB	Oberspreewald-Lausitz	51.4653	13.8680

Figure 1: data-frame browsing the records of Geo data of German cities areas

3 Exploratory Data Analysis

Before applying our approach, we have done some data analysis to obtain some general statistics. According to our analysis, Bavaria (Bayern) is the largest state with more than 2k postal codes 2. The state of North Rhine-Westphalia is the third largest state with more than 800 postal codes. We choose two cities in this state to explore for their further information. Bonn is a city on the Rhine banks with a population of over 300,000. About 24 km (15 mi) south-southeast of Cologne. It was the capital city of West Germany until 1990. The city Cologne(Koeln) is the largest North Rhine-Westphalia city and the fourth-most populous city in Germany. Cologne is the largest city on the Rhine, with slightly over a million inhabitants (1.09 million) within its city boundaries.

The bar plot 3 compare the number of categorized venues in these two cities sorted by Bonn. Like the city, Cologne is larger than Bonn; the number of venues in almost every category is also more significant. As you can see in the plot, 4, not all Bonn areas have the same number of venues. Most venues are in the city center (53111) and another (53125) region.

5 illustrates each category in the city Bonn and Cologne. As you see, they have very similar characteristics in this concern. In both, they are lots of cafe and restaurant. The Italian restaurants seem to be very common in both cities. Bonn has significantly fewer Gym and fitness centers than Cologne, although it has more nightclubs.

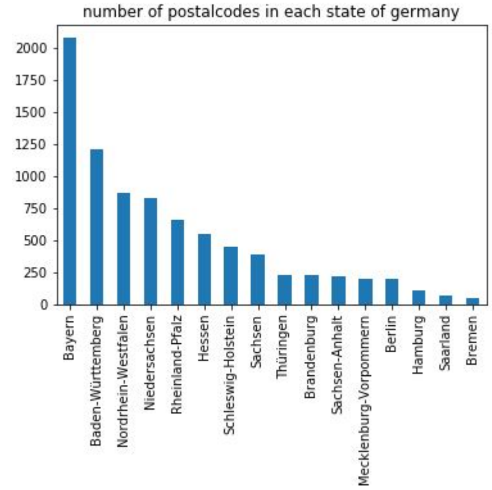


Figure 2: The number of postal codes in each state in Germany

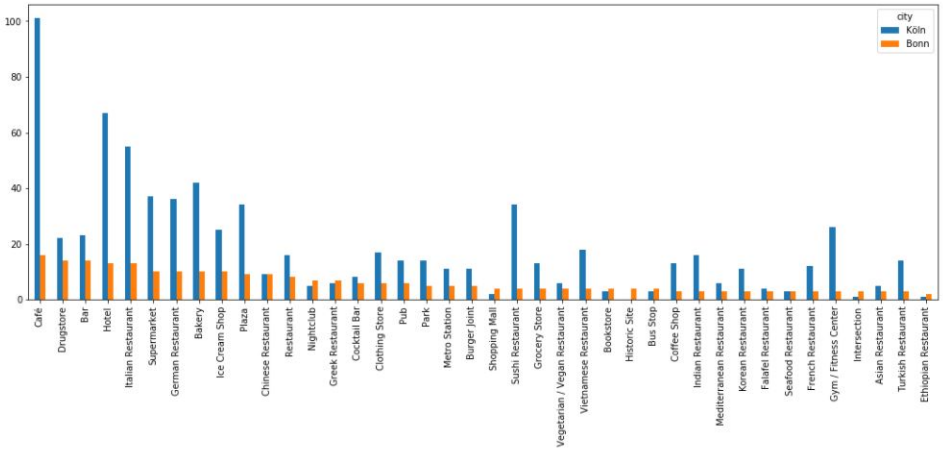


Figure 3: venue categories in cities Bonn and Cologne (Koeln)

3.1 Methodology and Feature selection

We use the machine learning method kmean to cluster Bonn's city areas to know a similar region. Then we Cluster all the regions of both city Bonn and cologne. Finally, we repeat the clustering for ten cities by averaging all venues in each city.

In the first step, we have collected the required data, including the postal codes of cities in Germany and geo-coordinates. After cleaning the data, we

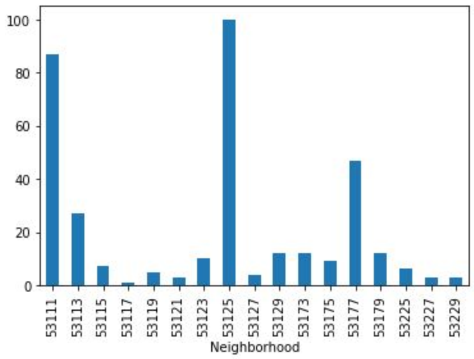


Figure 4: Number of venues groups in the Bonn city areas

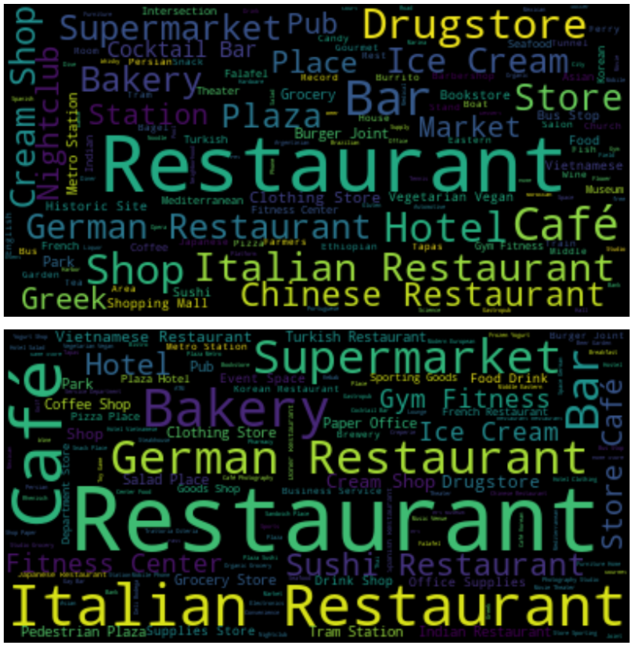


Figure 5: comparing the venues types and size in Bonn (top) and Cologne (down)

explore each Area based on its postal code to find the maximum of 100 venues in a 500-meter radius. Having the venues, we cluster the areas to see the similarities between areas (see figure 6).

In the next step, we compared the Area of one city with another by clustering all the two cities' Areas. We examined Bonn and Cologne's city areas to find similar regions in both cities (see figure 7).

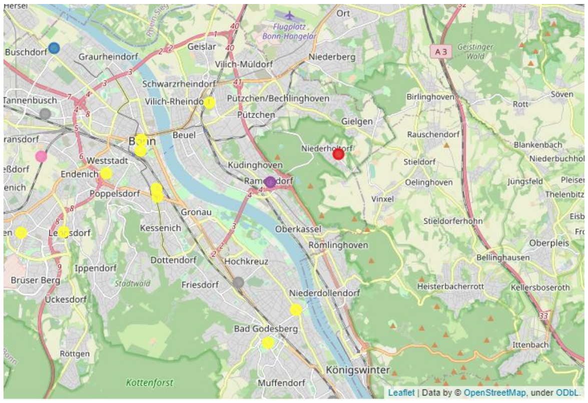


Figure 6: Clustering the Area of city Bonn

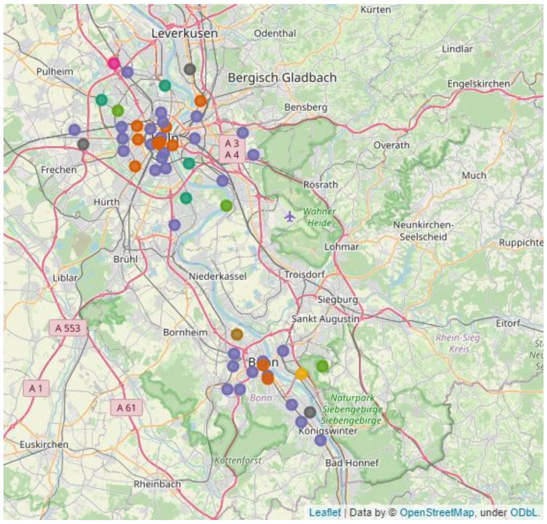


Figure 7: Clustering the Areas of both city Bonn and Cologne

The final part compares ten important cities in Germany, i.e., Hamburg, Dresden, Munich, Lubbock, Freiburg, Berlin, Potsdam, Trier, Dusseldorf, and Wiemar. We calculate the average of venue categories to embed the towns. We cluster the cities to get the groups of cities with similar features based on their venues' features (see figure 8).

4 Result and Conclusion

Our Analyse shows that Germany's city regions and cities have very similar forms based on their venues. Although they are some area with the very diverse

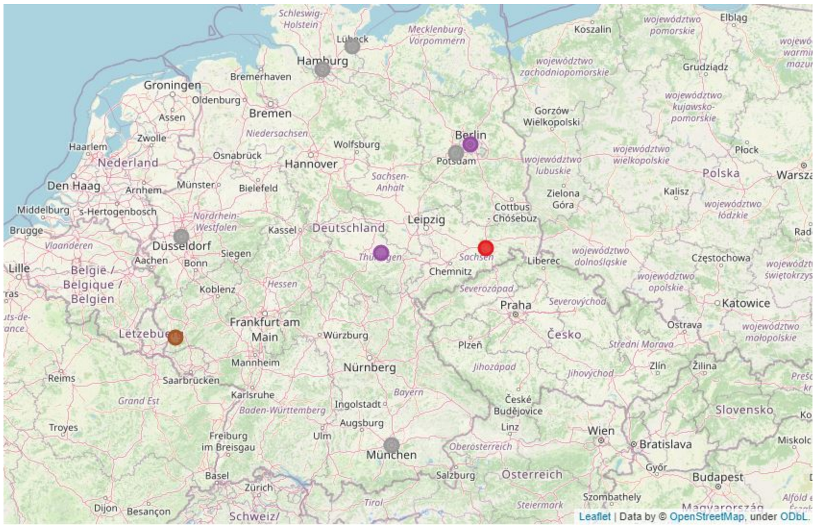


Figure 8: clustering ten cities in Germany according to their venues

feature, most of the area around the city centers have the same characteristics in the case of city regions. We also figured out that the site beside the train stations has very similar features.

When we compared the two cities' area, we figured out that the city centers' regions are typically similar. We also found some places which are not in the city center but still have the same categories.

Comparing different cities, we also get the same result. The towns differ maybe slightly. Since approaching ten towns, we have seven cities in one category and 2 in another, and there is one city that remains alone.

According to our analysis, we can say Berlin and Wiemar are very similar since they were always in the same categories with various clustering parameters.

This project aims to identify the places in different cities with similar characteristics based on their venues. We could recommend places when somebody like an area in a town and wants to find a similar position in another city. So if one wants to move from Bonn to cologne without knowing the cologne areas, we can offer some recommendations.

To get the perfect solution to the problem and cluster the cities, we also have to collect other resources and data. We can also consider additional features like the city size and population or near a river or mountain to compare the cities.

Finally, We would also need some gold-standard to evaluate how good is out clustering.