Final report:
German cities' Foursquare data comparability

Table of Contents

Introduction

Foursquare relies on users to add, describe, rate,... places.

It is not a widely used service in Germany. Therefore, probably not every city in Germany as an adequate large database which one can use to compare it to other cities.

This project will assess the properties of German cities with a population of > 100,000 people and their respective Foursquare venues data to give insights into the following questions:

- How much and what type of Foursquare Places data can a data scientist use to compare German cities?
- What are the limits of this comparison?
- Can one cluster the cities to enable data scientist to decide which cities are better suited for comparison than others? In which properties do the city clusters differ?
- Can one select cities that should be targeted with an ad campaign to promote the usage of Foursquare to expand the database?

The two stakeholders this data science project will target are:

- the "data scientist"
    - to decide whether he/she wants to use Foursquare for comparing cities in Germany
    - ... or to limit his/her comparison to (clusters of) cities with specific properties (population size, location, venues,...)
- the "Foursquare advertising team"
    - to select cities in which to promote the usage of Foursquare

Data

Data sources

This project will make use of different data sources, which are combined into one dataframe.

The list with German (large) cities from Wikipedia is the basis for the analysis:

https://de.wikipedia.org/wiki/Liste_der_Großstädte_in_Deutschland#Tabelle

It includes all German cities with a population of > 100,000 (definition of "Großstadt").
This project will use the city name, the population data from 2018, the area (sqkm), and the federal state.

To have some more information about the population of the cities the age distribution will be added from the German statistics website:

https://www.statistikportal.de/en/node/132

A difference in age groups may be one factor leading to different usage of Foursquare.
Due to complexity, the age distribution will be simplified into two groups: "<40 years" and ">40 years".
Although the age groups are only broken down to the state level and not the city level, they will be taken as representation for the cities.

Additionally, Geojson information for the German federal states from OpenDataLab.de will be used for a choropleth map to display the difference in age groups per state:

http://opendatalab.de/projects/geojson-utilities/

Using the city's name the geographic location will be retrieved with the help of Geopy / OpenStreetMap. The location data will be used to show the cities on a map (Folium) and, of course, to retrieve the venues using the Foursquare API.

Finally, the venues for each city from Foursquare API will be added using their geo-locations.

The Foursquare API restrictions limit the data that can be used for the project. E.g. the number of venues that can be retrieved per call per location is limited to 100.

The search radius will also have an impact on the number of venues. For simplification, this parameter will be fix. The radius is calculated from the median size (area in sqkm) of the cities.

Data exploration

First of all, I plotted various parameters against each other.

The fig. 1 shows the number of venues vs. the population size.

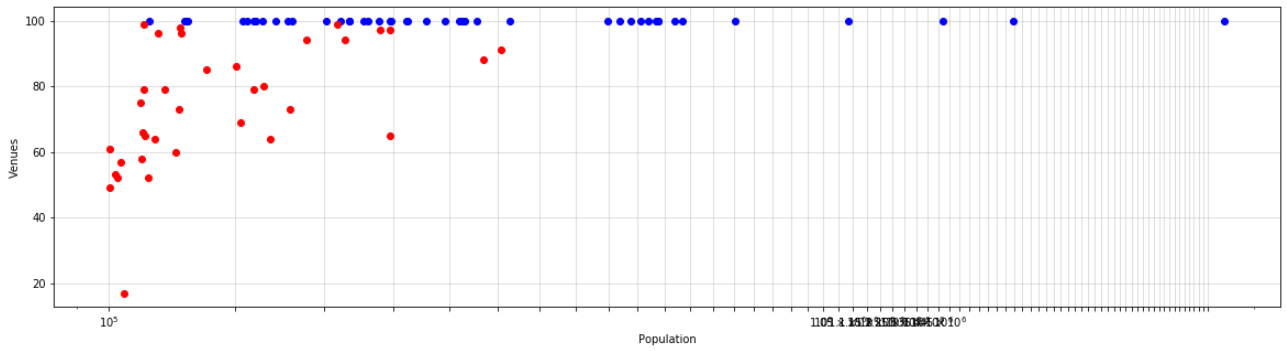

Figure 1

Germany has a few very large cities with more than 1,000,000 inhabitants. The largest city is Berlin with ~3,5 million inhabitants. This makes it quite difficult to plot, because one would not be able to tell the smaller cities apart on a decimal scale. Therefore, I chose a logarithmic scale for the population size.

The cities, which reach the Foursquare API call limit of 100 venues are displayed in blue. Every city with > 250,000 inhabitants reaches the limit. But also many cities between 100,000 and 250,000 inhabitants, too.

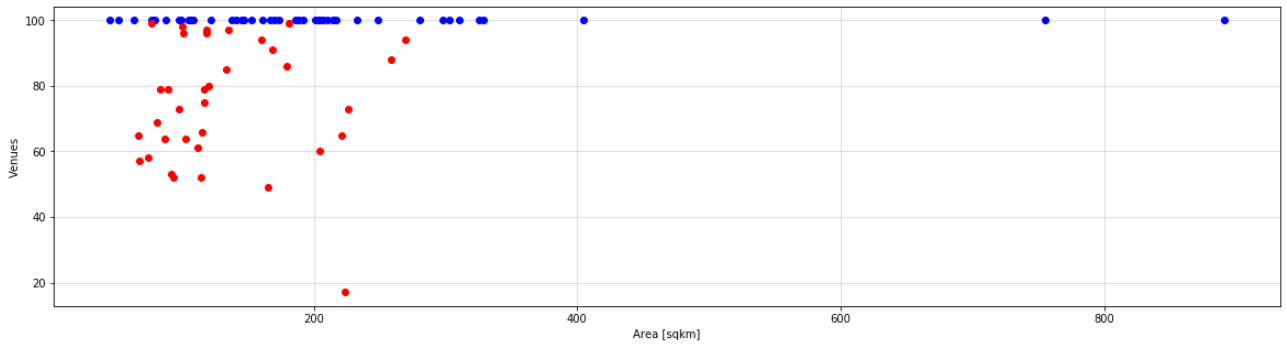

Figure 2

If you plot the number of venues vs. the area, fig. 2, the difference between cities with 100 venues and <100 venues becomes even smaller. Most of the cities range between an area of ~50 sqkm to ~300 sqkm and even the smallest city reaches the limit of 100 venues, whereas some large cities don’t.

A better parameter could be a ‘venue density’, as it standardizes the venues number and makes it more comparable between the cities of different sizes. Two ‘densities’ were created by dividing the number of venues per 100,000 inhabitants, and by area (sqkm).

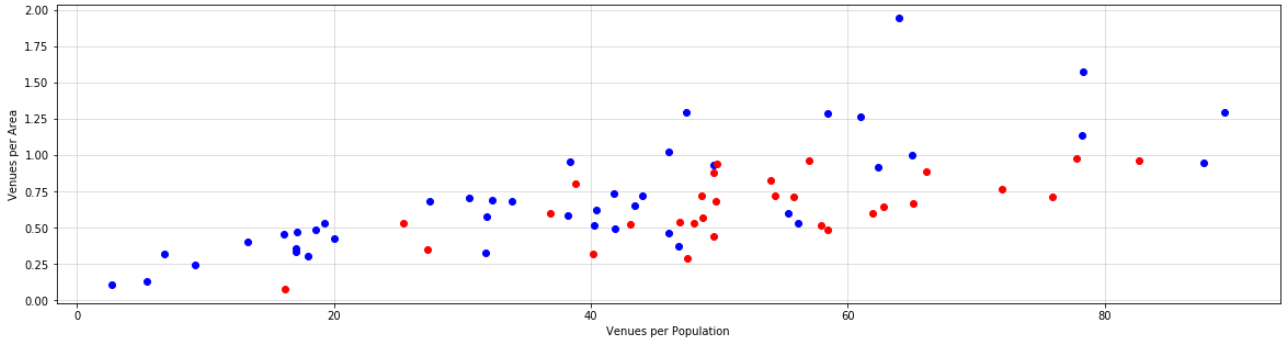

Figure 3

Plotting these two parameters against each other to get a quick impression shows that both groups are more or less scattered in the same region (fig. 3). This means, even with these standardized parameters, one cannot tell the groups apart easily.

The last figure 4 plots the 'Venues per Population' parameter against the share of people younger than 40 years of age of the total population of the federal state (PSA).

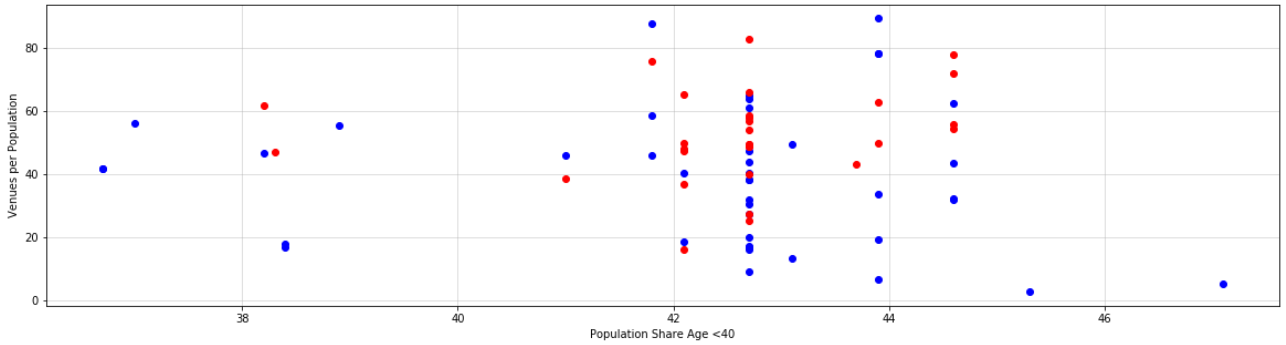

Figure 4

The first intuition was that cities with a low PSA might show a smaller venue density, because the more young people the more would Foursquare been used. This intuition proves wrong, because there are even more cities with a low PSA (< 40%) that reach the limit of 100 venues than there are cities that have less than 100 venues.

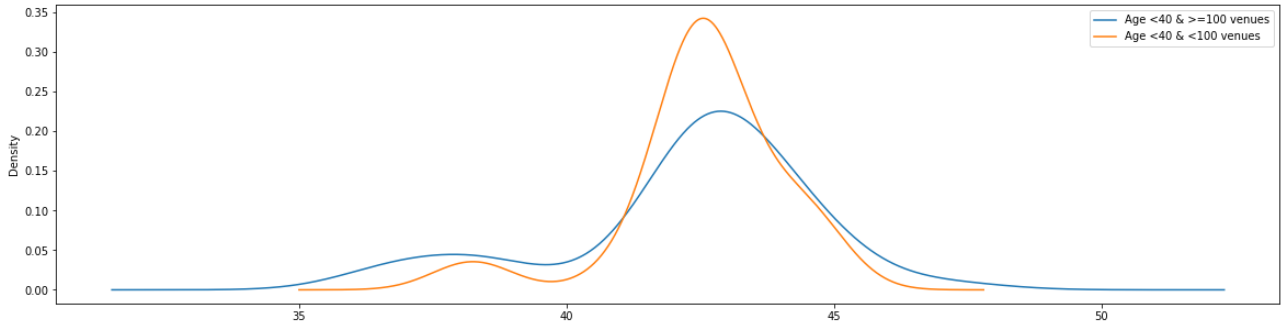

Figure 5

A density plot (KDE) of the shows this even better (fig. 5). The blue curve shows the density (derived from the number of cities per PSA) of the cities with 100 venues, the orange curve the density for cities with <100 venues. There a slightly more cities with 100 venues in the low PSA region than cities with <100 venues, which one can derive from the higher 'blue bump' at around 38%.

Fig. 6 shows the location of the cities (blue: 100 venues, red: <100 venues).

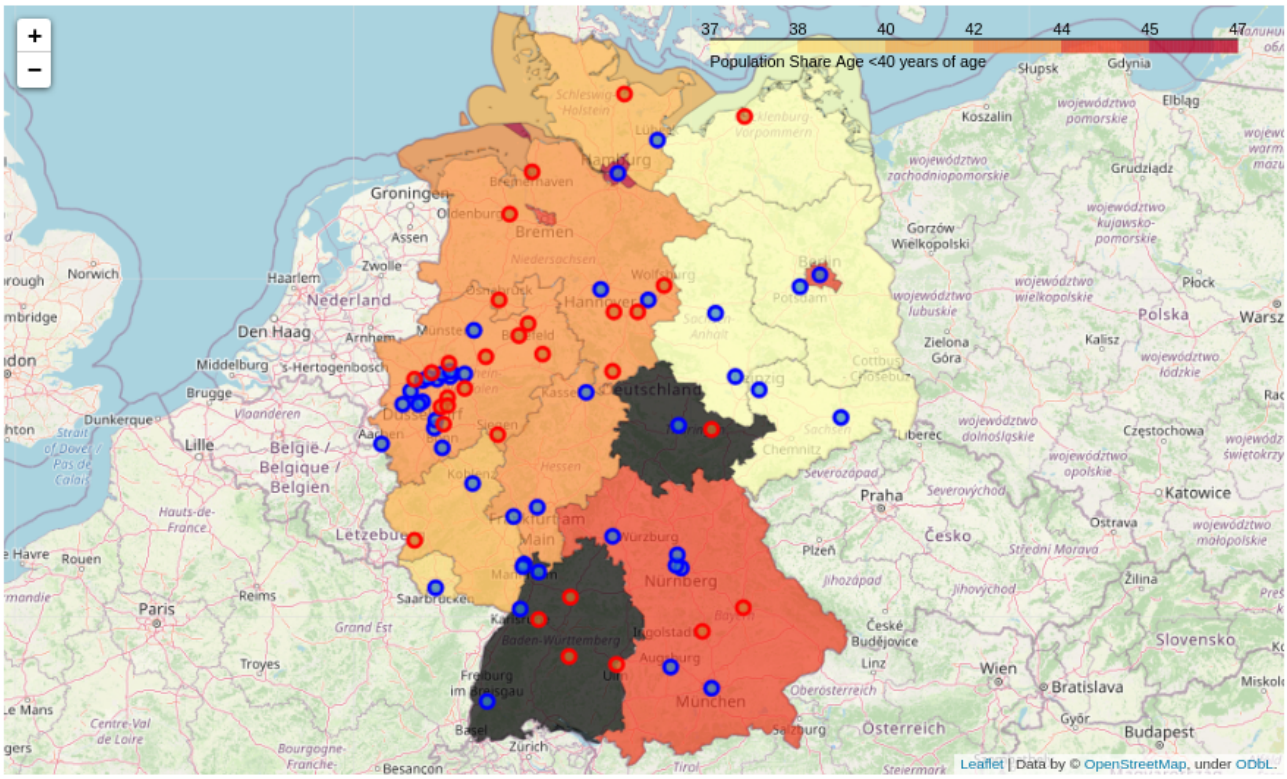

Figure 6

Methodology

In the first step, we created a dataframe with all big cities in Germany, their properties, their location and their Foursquare venues.

The limit of 100 venues per city per Foursquare API call distorts results while comparing he cities. Therefore, all cities with a venue count of 100 will not be included in the dataframe for further analysis. The reduced dataframe of cities with < 100 venues consists of 35 cities.

In the second step, the 35 cities will be clustered into a 'manual' cluster of 4 equal-sized groups by quartiles of the parameter 'Population' (count) and also into 4 clusters by using the K-Means ML method.

Lastly, the results of these two clusters will be compared regarding the cities properties and the top venues (limited to 10 cities and the top five venues).

Results

Manual clustering

First of all, we calculate the quartiles for each data column.

	Population	Area	Population Share Age <40	Venues
0.25	111546.75	92.590	42.1	59.75
0.50	124846.50	114.725	42.7	75.00
0.75	163401.00	153.680	42.7	89.50
1.00	354382.00	258.820	44.6	98.00

Figure 7

Fig. 7 shows the table with the quartiles of 25%, 50%, 75%, and 100%. The quartiles of the population column will be used to split the cities into four groups and label the accordingly. Fig. 8 shows the result (limited to the first five rows) with the new column ‘Quartile’.

	Quartile	City	Population	Area	Latitude	Longitude	State	Population Share Age <40	Venues	Venues per Population	Venues per Area	1st Most Common Venue	2nd Most Common Venue	3rd Mos Commo Venu
5	3	Regensburg	152610	80.70	49.019533	12.097487	Bayern	43.9	76.0	49.800144	0.941760	German Restaurant	Hotel	Supermark
6	3	Ingolstadt	136981	133.37	48.763016	11.425040	Bayern	43.9	86.0	62.782430	0.644823	Supermarket	German Restaurant	Italia Restaura
16	4	Wuppertal	354382	168.39	51.264018	7.178037	Nordrhein-Westfalen	42.7	90.0	25.396324	0.534474	Supermarket	Café	Ba
17	4	Bielefeld	333786	258.82	52.019101	8.531007	Nordrhein-Westfalen	42.7	91.0	27.262977	0.351596	Supermarket	Hotel	Train Statio
25	4	Hagen	188814	160.45	51.358294	7.473296	Nordrhein-Westfalen	42.7	92.0	48.725200	0.573387	Supermarket	Bakery	Clothin Stor

Figure 8

K-Means clustering

This manual splitting is being compared to a K-Means clustering of the cities. For a good comparability, also four clusters will be used to fit the data.

The data used for K-means consists of the columns Population, Area, PSA and Venues (see Fig. 9).

	Population	Area	Population Share Age <40	Venues
5	152610	80.70	43.9	76.0
6	136981	133.37	43.9	86.0
16	354382	168.39	42.7	90.0
17	333786	258.82	42.7	91.0
25	188814	160.45	42.7	92.0

Figure 9

Comparison of clustering methods

The results are compared using boxplots, plotting each cluster as an invidual boxplot.

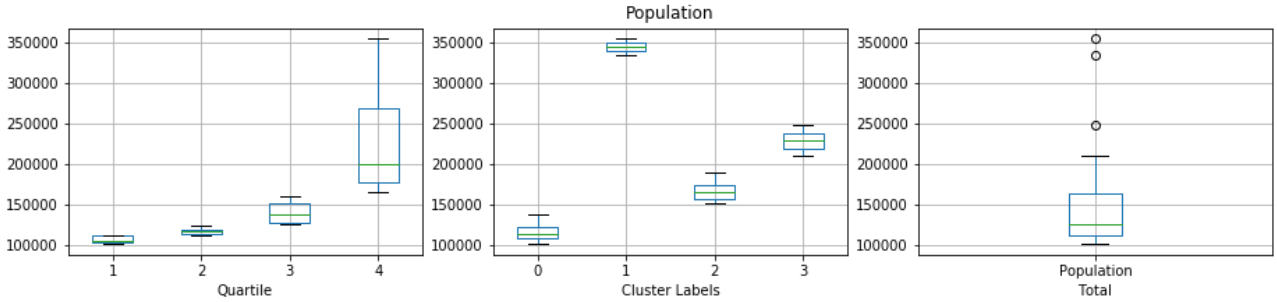

Figure 10

The biggest accordance and also difference between the manual clustering and the k-means clustering is by the population data. Both methods have the statistically best split between the groups within the population size, as the boxplots for the groups of each method do not overlap at all. This comes as no surprise for the manual method, as we deliberately split the groups by quartiles of the population size. The K-means on the other hand seems to use the population size as the most important factor for the split, but does not split the cities into equal sized groups.

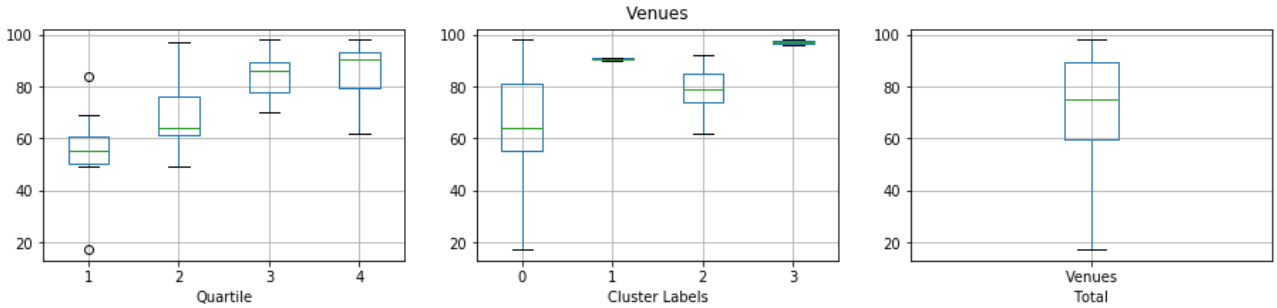

Figure 11

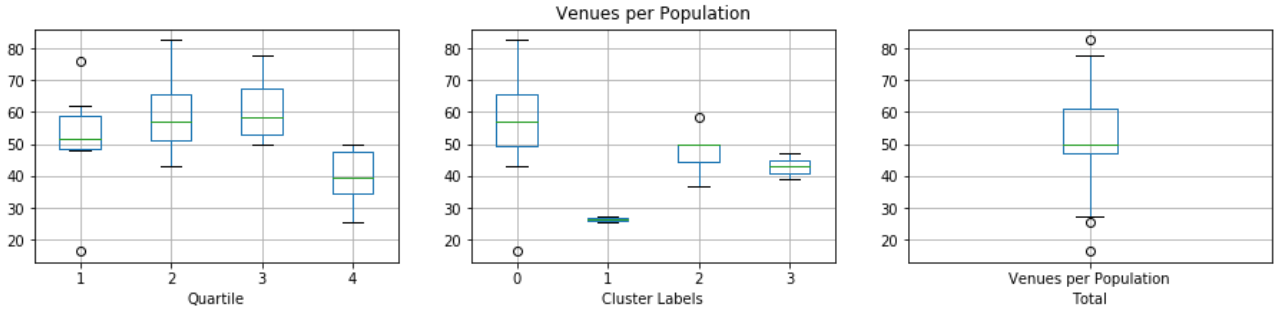

Figure 12

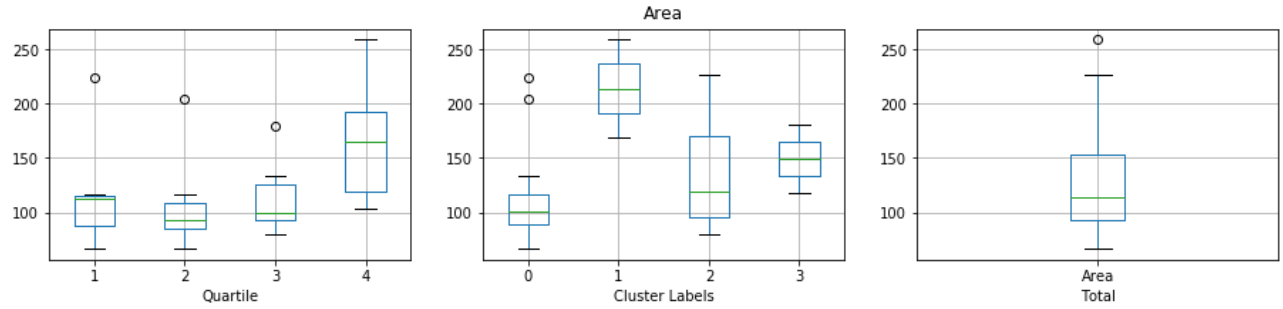

Figure 13

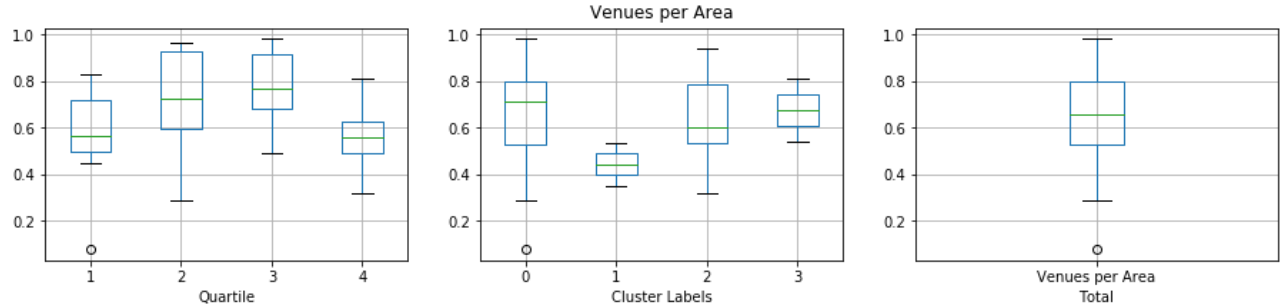

Figure 14

The medians of Venues per Population (Fig. 12) and Venues per Area (Fig. 14) data are lowest for the largest cities in both k-means cluster 2 and population quartile 4. The K-means cluster 2 sets itself apart from the other clusters even better than quartile 4 from the quartile groups.

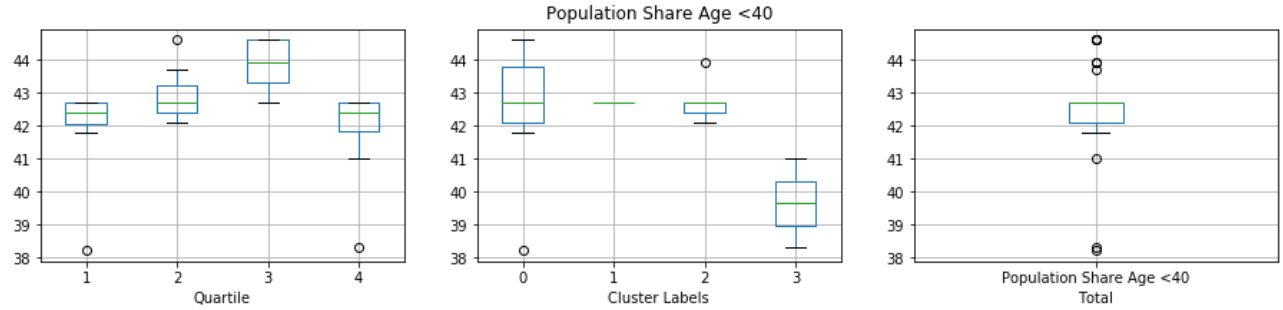

Figure 15

The Age Share <40 of quartile 4 / cluster 2 (Fig. 15) does not differ significantly to at least on other group, respectively the median of all cities combined.

Comparison of the top venues of the K-means clustered data

The comparison of the venue categories will be limited to the top 5 venues and the top 5 cities of the K-means cluster 0 and both cities of cluster 2.

	Cluster Labels	City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
6	0	Ingolstadt	Supermarket	German Restaurant	Italian Restaurant	Café	Drugstore
32	0	Bottrop	Supermarket	Hotel	Clothing Store	Café	Drugstore
33	0	Recklinghausen	Supermarket	Drugstore	Café	Clothing Store	Italian Restaurant
34	0	Bergisch Gladbach	Supermarket	Drugstore	Hotel	Italian Restaurant	Hardware Store
35	0	Remscheid	Supermarket	Gas Station	Café	German Restaurant	Drugstore
5	2	Regensburg	German Restaurant	Hotel	Supermarket	Café	Plaza
25	2	Hagen	Supermarket	Bakery	Clothing Store	Hotel	Café
26	2	Hamm	Supermarket	Drugstore	Ice Cream Shop	Bakery	Big Box Store
28	2	Solingen	Supermarket	Café	German Restaurant	Drugstore	Restaurant
31	2	Paderborn	Supermarket	German Restaurant	Hotel	Italian Restaurant	Café

Figure 16

Fig. 15 shows that the supermarket category dominates the 1st most common venues. Apart from that, one can note that in group 'cluster label 0' more 'store-like' venues (Drugstore, Clothing Store, Hardware Store,...) are represented than in group 'cluster label 2'.

On the other hand, in group 'cluster label 2' more 'travel-related' and 'touristic' venues like hotels and cafés dominated the 2nd to 5th most common venues.

Map showing the cities of the two clustering methods

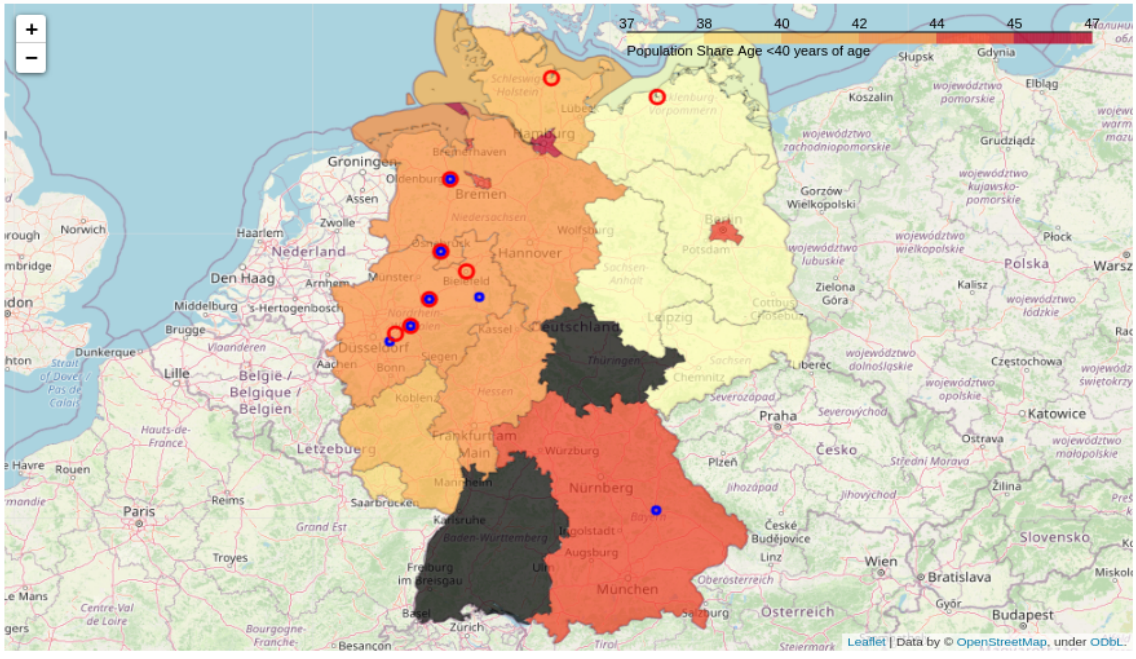

Figure 17: Red circles: quartile method cluster #4; blue dots: K-Means cluster #2

Results and Discussion

The analysis shows that 'manual' clustering by 'gut' using the most obvious data columns and a splitting into quartiles is not a necessarily useless first approach when compared to the K-means ML method.

Plotting the different clusters against each other and for the different variables, I chose the parameter 'Venues per Population' as the significant parameter for identifying the cities that have the biggest potential in expanding the Foursquare user base e.g. by advertising the usage of Foursquare.

The resulting dataframe of cities with the 25% largest population size (quartile 4) includes 9 cities. These 9 cities are located mainly in the northern half of Germany and in the western states as well as in the eastern federal states.

When using the K-means method, cluster 2 with only two cities sticks out. This cluster (label 2) includes the cities Bielefeld and Wuppertal, which have large population but also a low count of venues. Thus resulting in very low 'Venues per Population' values.

The 'manually' grouped cluster 'quartile 4' includes these cities as well. But the K-means method helped to narrow down the group of cities to be targeted by advertisement to 2 instead of 9 cities.

When comparing the type of venues of the cities Wuppertal and Bielefeld to the K-means cluster 0, which includes the highest 'Venues per Population' values, one notes that no 'stores' are represented in the top 5 venues in Wuppertal and Bielefeld.

Conclusion

Concluding this project I'd like to look at the questions that were stated in the business problem chapter and answer them:

- **How much and what type of Foursquare Places data can a data scientist use to compare German cities? What are the limits of this comparison?**
  - When using Foursquare's API the 'common' data scientist is limited to 100 venues per API call. This limit may be bypassed by using a smaller search radius and multiple searches per city.
  - Apart from that, the usual variety of Foursquare data can be used. In this project, I limited myself to the venue (category) data.
  - The age share data probably does not necessarily represent the age structure of the cities, as it was compiled on federal state level and not on city level.

- **Can one cluster the cities to enable data scientist to decide which cities are better suited for comparison than others? In which properties do the city clusters differ?**
  - Probably all cities which are limited by the Foursquare API call to 100 venues have a large enough Foursquare database to be used for comparison as usual.
  - If you focus on the cities with less than 100 venues per API call, the largest (by population) of these cities tend to have a low venue per population count. These cities probably have a small Foursquare database compared to the potential user base.
  - The cities of cluster 2, Wuppertal and Bielefeld, don't have any 'store'-like venues in their top 5 venues.

- **Can one select cities that should be targeted with an ad campaign to promote the usage of Foursquare to expand the database?**
  - The cities of cluster 2, Wuppertal and Bielefeld, should be targeted, because they have quite a large population (>300,000 inhabitants) and a high share of young people, which, to me, are the factors for a large potential Foursquare user base.