Trip Advisor: Clustering German Cities by Venues

by Tobias (https://github.com/TobiasGuggemos), May 2020.

1. **Introduction**:
   This project is about German cities. The goal is to support decisions about the next spot for a city trip. Potential stakeholders are all people interested in city trips in Germany. We will cluster the cities based on the venues near to the city centre. This will help to choose an appropriate spot for the next city depending on the purpose of the trip and the interest in different venue types. The city clusters can be used to find similar cities to already visited and much appreciated cities and trips.

2. **Data**:
   First, we want to identify the largest cities in Germany. Therefore we will use a list on wikipedia of large german cities, large defined as: with a population of min 100k. We also need to have all the city latitude and longitude coordinates which we can explore with Geopy, a Geocoding libraries for Python.
   An extract of the dataframe with the cities and their geo coordinates as well as the cities plotted on a map are displayed below:

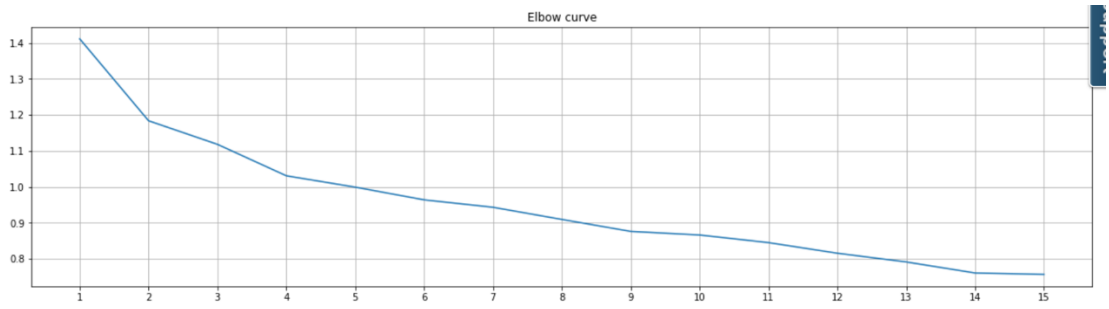
	city	latitude	longitude
0	Berlin	52.517037	13.388860
1	Hamburg	53.543764	10.009913
2	Munich	48.137108	11.575382
3	Cologne	50.938361	6.959974
4	Frankfurt am Main	50.110644	8.682092
...	...	...	...
74	Erlangen	49.598119	11.003645
75	Moers	51.451283	6.628430
76	Siegen	50.874980	8.022723
77	Hildesheim	52.152164	9.951305
78	Salzgitter	52.150372	10.359315

For all of these cities we need to list the venues around the city center. We will use Foursquare API and Get Venue Recommendations for a given place. Furthermore, let's define the number of requested venues per city and test/start with 100. We consider venues within a circle with a radius of 2,000m around the city center. For the city clustering we also need to know the venue category. This data is also collected via Foursquare.

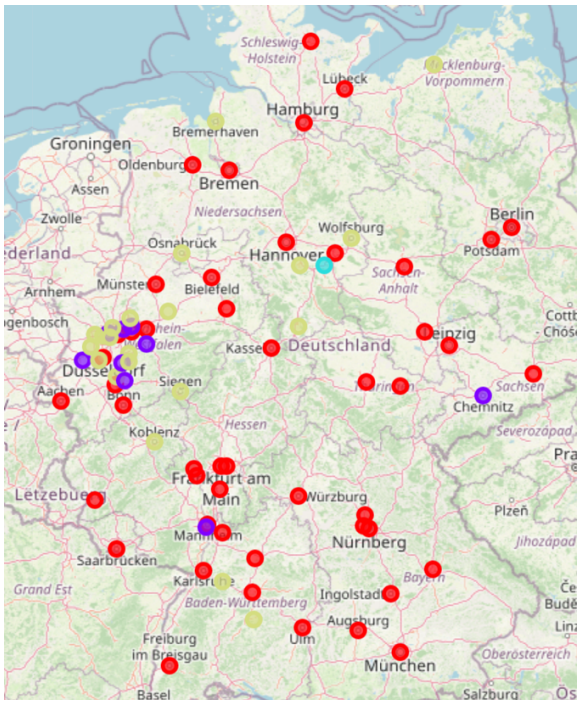
To prepare the data for the cluster analysis, we use One Hot Encoding. In doing so, we create dummy variables for the column 'venue category' and add the column ‘city’ to this dataframe. Next, we group this dataframe by ‘city’ and calculate the mean for all the venue categories. Finally, we display the top 10 venue categories per city:

	city	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Aachen	Bakery	Park	Café	German Restaurant	Supermarket	Plaza	Bar	Italian Restaurant	Hotel	Coffee Shop
1	Augsburg	Italian Restaurant	Café	German Restaurant	Hotel	Steakhouse	Bar	Drugstore	Burger Joint	Brewery	Beer Garden
2	Bergisch Gladbach	Supermarket	Drugstore	Café	Bakery	Clothing Store	Shopping Mall	Wine Shop	Gym / Fitness Center	Museum	Bank
3	Berlin	Coffee Shop	History Museum	Hotel	Monument / Landmark	Plaza	Bookstore	Concert Hall	Historic Site	Gourmet Shop	Drugstore
4	Bielefeld	Bar	Middle Eastern Restaurant	Hotel	Supermarket	Greek Restaurant	Café	Italian Restaurant	Asian Restaurant	Restaurant	Nightclub
...	...	...	...	...	...	...	...	...	...	...	...

3. **Methodology**:
   To support travel decisions we aim at a clustering of the cities. We start with the K-Means algorithm since it is most popular for clustering in general, easy to understand and implement as well as efficient also with large data sets (even though we do not handle large data here). For the decision about the number of clusters we draw an elbow curve. As evaluation metric we use inertia, the sum of intracluster distances (within-cluster sum-of-squares). Inertia can be regarded as a criterion of how internally coherent clusters are. We have also analyzed the mean of the Euclidean distances from all data points to corresponding cluster centroids with very similar results. We should mention that Inertia assumes convex and isotropic clusters and euclidean distances tend to become inflated in high-dimensional spaces (see https://scikit-learn.org/stable/modules/clustering.html). For details please go to scikit-learn.org.
   Later on, we will also conduct hierarchical clustering for verification of our results.
   A clear elbow is visible at k=2. Since we want more and smaller city clusters, we start our analysis with k=4 clusters in the first run.



4. **Results**:
   The resulting clusters are shown in different colors in following map:



Extract Cluster 0:

	city	latitude	longitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Berlin	52.517037	13.388860	Coffee Shop	History Museum	Hotel	Monument / Landmark	Plaza
1	Hamburg	53.543764	10.009913	Hotel	Coffee Shop	Art Gallery	Plaza	Italian Restaurant
2	Munich	48.137108	11.575382	Café	Plaza	German Restaurant	Ice Cream Shop	Cocktail Bar
3	Cologne	50.938361	6.959974	Hotel	Plaza	Coffee Shop	Italian Restaurant	Café
4	Frankfurt am Main	50.110644	8.682092	Café	Park	Bar	Plaza	Burger Joint
5	Stuttgart	48.778449	9.180013	Cocktail Bar	Plaza	Sushi Restaurant	German Restaurant	Bar
6	Düsseldorf	51.225402	6.776314	Coffee Shop	Café	Japanese Restaurant	Hotel	Park
7	Dortmund	51.514227	7.465279	Café	Coffee Shop	Italian Restaurant	Pub	Ice Cream Shop

Extract Cluster 1:

	city	latitude	longitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
24	Gelsenkirchen	51.511032	7.096012	Supermarket	Drugstore	Park	Bus Stop	Bakery
25	Mönchengladbach	51.194698	6.435364	Supermarket	Café	Gym / Fitness Center	Clothing Store	German Restaurant
27	Chemnitz	50.832261	12.925298	Supermarket	Drugstore	Asian Restaurant	Nightclub	Café
35	Oberhausen	51.469614	6.851444	Supermarket	Drugstore	Gym / Fitness Center	Music Venue	Hotel
40	Hagen	51.358294	7.473296	Supermarket	Bakery	Clothing Store	Italian Restaurant	Doner Restaurant
45	Ludwigshafen am Rhein	49.470411	8.438157	Supermarket	Clothing Store	Hotel	Italian Restaurant	German Restaurant
49	Solingen	51.171247	7.083900	Supermarket	Café	Clothing Store	Hotel	Sandwich Place
51	Herne	51.538039	7.219985	Supermarket	German Restaurant	Drugstore	Bus Stop	Fast Food Restaurant
71	Bergisch Gladbach	50.992930	7.127738	Supermarket	Drugstore	Café	Bakery	Clothing Store

Extract Cluster 2:

	city	latitude	longitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
78	Salzgitter	52.150372	10.359315	Gym / Fitness Center	Bakery	Shopping Mall	Grocery Store	German Restaurant

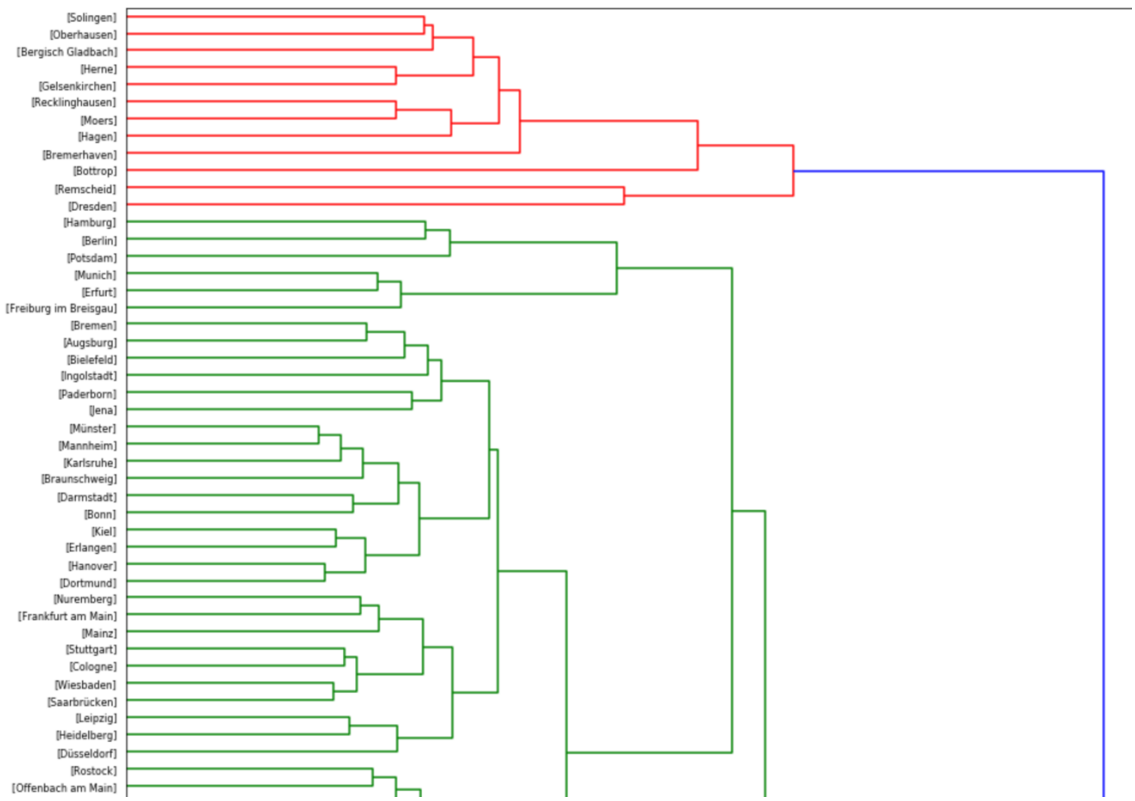
Extract Cluster 3:

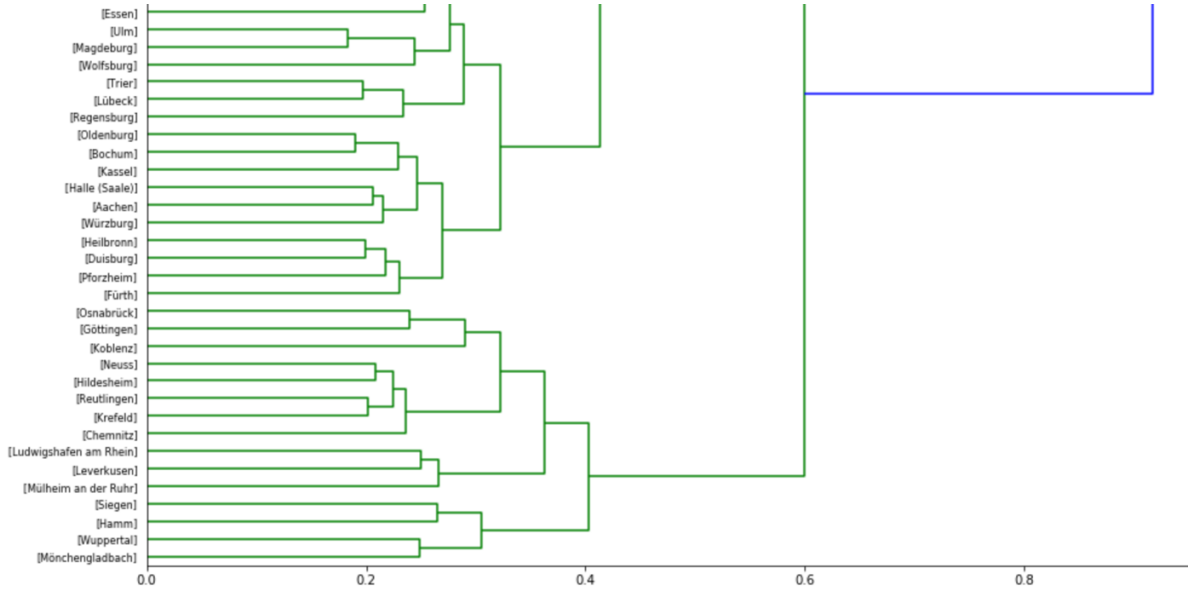
	city	latitude	longitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
14	Duisburg	51.434999	6.759562	Café	Supermarket	Italian Restaurant	Restaurant	Bakery
16	Wuppertal	51.264018	7.178037	Supermarket	Café	Clothing Store	Fast Food Restaurant	Plaza
33	Krefeld	51.333120	6.562334	Supermarket	Café	Drugstore	Clothing Store	Bakery
38	Rostock	54.092444	12.128613	Hotel	Supermarket	Restaurant	Café	Indian Restaurant
41	Hamm	51.680409	7.815197	Supermarket	Drugstore	Platform	Bakery	Fast Food Restaurant
43	Mülheim an der Ruhr	51.427293	6.882919	Supermarket	Hotel	Park	German Restaurant	Museum
47	Leverkusen	51.032474	6.988119	Hotel	Supermarket	Drugstore	Soccer Stadium	Fast Food Restaurant
48	Osnabrück	52.266837	8.049741	Café	Supermarket	Bar	Nightclub	Asian Restaurant
52	Neuss	51.198178	6.691648	Clothing Store	Café	Supermarket	Drugstore	Bakery
59	Wolfsburg	52.420559	10.786168	Hotel	Italian Restaurant	Drugstore	Supermarket	Café
63	Pforzheim	48.890885	8.702953	Café	Supermarket	German Restaurant	Italian Restaurant	Clothing Store
64	Göttingen	51.532760	9.935205	Supermarket	Café	Italian Restaurant	Drugstore	Hotel

Cluster 2 seems to be an outlier including only one city. The city of Salzgitter is the smallest of the whole data set. Furthermore, the Foursquare API call delivered only very few venues for this city. Since K-Means is sensitive to outliers (https://www.researchgate.net/publication/293061584_Comparative_Study_of_K-Means_and_Hierarchical_Clustering_Techniques),
we exclude this city for our further analysis. For verification, we also applied hierarchical clustering. Again, Salzgitter represents an own cluster.

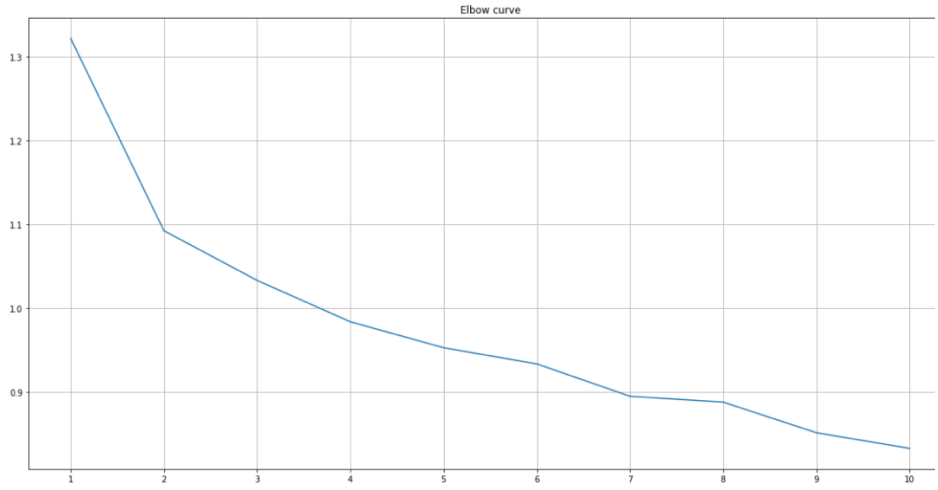
We're quite happy with the results of the other clusters: Scanning the cities of each cluster and their most common venues, we identify that Cluster 0 seems to be dominated by Cafes/Coffee shops (followed by Hotel and Plaza), Cluster 1 includes a lot of supermarkets and drugstores () and Cluster 3 seems to be somewhere in the middle between Cluster 0 and 1 with Supermarkets, Cafes and Hotels.

Running the analysis again and without the outlier city, the results of hierarchical clustering are:

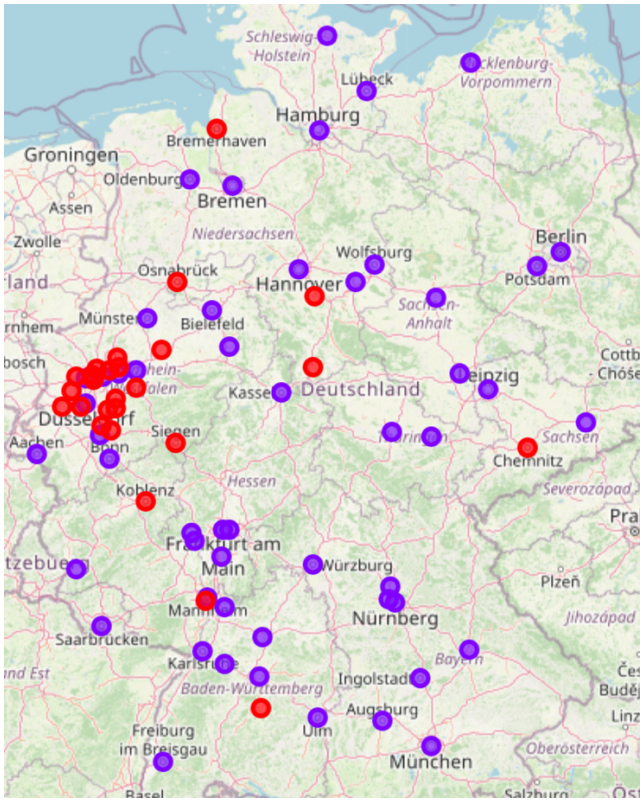




We can quickly recognize 2 main clusters. The first cluster is colored red and including many cities with supermarkets as most common venue type, the second one in green with a lot of “café cities”.
We have also applied the K-Means algorithm excluding the smallest city in the dataset. The update elbow curve:



We see again a strong kink/break at k=2. Running K-Means with k=2 delivers following results:



The first cluster includes the cities with red markers. The top venues are:

```
Café
Hotel
German Restaurant
Italian Restaurant
Plaza
```

The cities of the second cluster are colored in purple on the map above. Their most common venues are:

```
Supermarket
Café
Drugstore
Hotel
Clothing Store
```

5. **Discussion**:
   The results clearly point out that there are two heterogeneous clusters. Repeated analyses indicate quite robust results. The same applies for different methods: Hierarchical clustering as well as K-Means with k=2 steadily separate the data points into “supermarket” vs. “cafe” city centers. The break in the elbow curve at k=2 refers to this finding. Further clusters seem to be difficult to interpret. For that reason we finish the analysis with a focus on two clusters. Nevertheless, it is very useful to have the city hierarchy and to have a look at the sub clusters. Users can e.g. search for favorite cities and identify similar cities for the next trip.
   For our stakeholders we propose to focus on the purple cluster of the city map (clustered by K-Means) and the green colored cities in the city hierarchy (hierarchical clustering). Both clusters represent cities with cafes and restaurants as most common venues in the city center and strongly overlap. Since our stakeholders are coffee lovers and also look for a city with a large selection of good restaurants, this cluster seems to be a good starting point to find great locations for the next trips. Therefore we schedule a meeting for further discussion – especially on the list of cities of the favored cluster.

6. **Conclusion**:
   The clustering of Germany city centers discovers two significantly different clusters. As a result we could narrow down the number of potential cities for the next journey. For our stakeholders, we have identified a cluster with cities that are characterized by a lot of cafes and restaurants nearby the city center. For further insights it would be interesting to see how more criteria besides venues affect the results.