# Where to visit to in Germany

# Table of Contents



## *1. Introduction*

### Background

I moved from London to Stuttgart a few years ago, for work reasons. At the time I knew very little about Stuttgart or Germany. I still know very little about other cities in Germany and would like to visit more cities to explore them.

I like running and find this an excellent way to experience other towns and cities. To fuel the activities, I like to eat lots. So I thought it would be interesting to see which cities provide good running opportunities as well as a variety of restaurants offering different international cuisine.

### The problem

Create a shortlist of cities in Germany to visit. The short-list should include cities that have a wide range of international cuisine and some good running and cycling opportunities.

### Target audience

With the growth of people taking 'active' holidays, especially cycling and running, this would be a good tool to explore places to visit in Germany.

## *2. Data*

### German Cities

The long-list of cities will come from the German definition of a Großstadt, which is one with a population of more than 100,000. There is a list of these cities on wikipedia.[1]

This data contains latitude and longitude but will need to be wrangled as it is presented as Degree Minutes Seconds (DMS)  pairs and not as decimal values

### Restaurants

Foursquare offers information about different venues around a location I will use this to identify restaurants with different types of cuisine, e.g Italian restaurants, Japanese restaurants etc. [2]

I will need to identify food venues and keep only these, discarding other venues such as museums or shops.

### Running segments

Strava is an App that allows users to upload sporting activities. Users can create 'segments' which are short stretches of running or cycling routes. Users compete over these segments. From the Strava API, information about the segments can be found such as the gradient and length. This will be used to identify good running opportunities.[3]

## *3. Methodology*

Jupyter Notebook was used for the analysis and GitHub as the repository.

The methodology used was unsupervised learning, by K-means clustering analysis, with the goal to group the different cities into clusters representing their mix of eating establishments and running routes. One possible drawback of this methodology is the relatively small sample size (79 cities) that I had. Cluster analysis is best performed on sample sizes of 200 or more.

Before performing any cluster analysis the data had to be wrangled and merged.

### Cities data

The table of German Cities was scraped from Wikipedia. This table included a long text string which had to be parsed to extract the latitude and longitude in decimal degree format. This was error-checked by mapping all points on a map. From here I could see that a small number of them were showing as outside of Germany. This was because the format of the string in the wikipedia table was inconsistent for these points, so I had to correct them manually.

### Running data

Strava API has an running segment explore endpoint. When provided with North-East and South-West coordinates this API returns the 10 most popular segments within the square

formed by these coordinates. I had to decide what size square to use in order to get a good representation of the types of running routes available within a city. In the end I decided that a square with sides of 10km would be about right. Converting this into degrees actually gave me a square 11km x 11km, although this varied on the actual latitude of the city.

Having calculated the boundary points for my city, I checked this for Stuttgart as I know this city, by mapping the boundaries to see how far this stretched. This looked to be a reasonable travel zone for someone staying in the city centre. I requested the run segments for Stuttgart and based on this data, had to decide how to categorise runs. I decided on two categorisations

- steepness – hilly segments were ones with an uphill or downhill gradient of >2%;
- length – long segments were ones >1km

I then extracted the segments for all of the 79 cities and produced a table showing for each city what proportion of runs were hilly vs flat and long vs short.

## Restaurant data

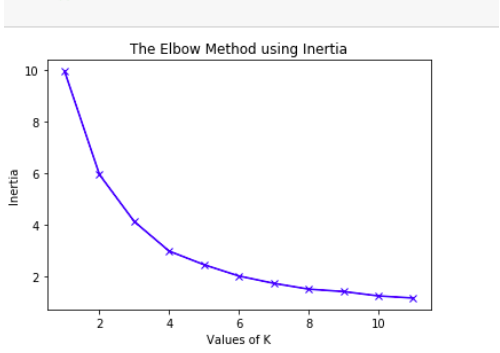
The foursquare API was used to explore venues in each of the 79 cities. (This was a constraint of the project definition).

I extracted the venue data for each of the cities, limiting results to 100 and using a 5km radius to cover a similar area to the running segments. I needed to identify only eateries, so I produced a list of all the venue categories with a count of entries. From here I could see that most eateries included the word 'Restaurant' in the venue category name, but there were also a few others such as Cafe, Burger Joint, Bistro. I produced a manual list of the additional eateries that I wanted to included in my list.

I then extracted just the eatery venues and created a table for each of the cities showing the proportion of different types of eating establishment in the city.

## Cluster Analysis

Finally I was ready to merge the running and restaurant data and perform cluster analysis. The first step was to determine how many (K) clusters to model. I used elbow method and inertia to test K values from 1 to 12.  As can be seen from the graph below, an elbow is identifiable at K=4.




I then ran my cluster analysis model on the 79 cities. For this I used the agglomerative

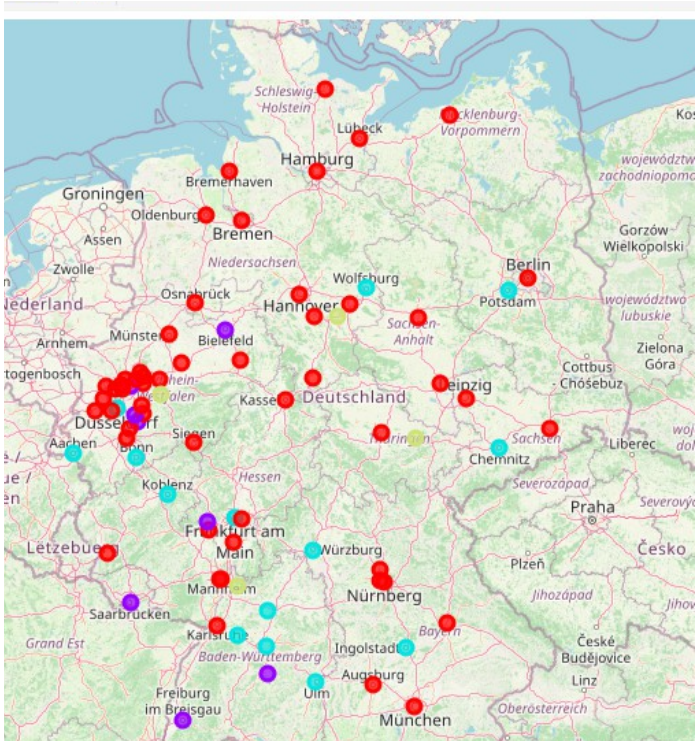
clustering and the added the cluster labels to the table of cities.

Finally I created a table for each city showing the 5 most popular eating establishments and the split of hilly/flat and long/short run segments. From this I was able to identify the characteristics for each cluster and give them descriptions.

The results were visualised using Folium mapping library.

## *4. Results*

The map below shows the result of the clustering and the descriptions of each cluster.




| Cluster | Description |
|---|---|
| 0 | lots of cafes, very flat running |
| 1 | German restaurants, mainly short runs |
| 2 | German and Italian restaurants, hilly and short runs |
| 3 | even mix of hilly/flat and long/short runs, assorted restaurants |

## *5. Discussion*

## Trip recommendations

As can be seen from the map and the descriptions, there are some good options for running and eating in Germany!

Stuttgart, where I live is hilly and is included in cluster 2, so if I wanted to visit other cities that have similar characteristics then I could look at the list of cluster 2 cities and pick one of these. As to be expected, most of these are located in the south of the country closer to mountains. However, if I wanted some different running, perhaps flatter then maybe some of the northern cities that tend to be in cluster 0 would be of interest.

## Improvements

In doing this analysis, I identified a number of improvements for future iterations.

1.) Hierarchical Analysis

Best practice suggests that Cluster Analysis is best performed on sample sizes of greater than 200. As I only had 79 cities in my sample, an alternative approach would be to use hierarchical analysis and see what results this produced.

2.) More Strava segments
The Strava API will only retrieve 10 segments. Online forums advise a workaround that requires breaking down the city into more (e.g 4 or 9) smaller sections and run the API for each of these. It is suggested that the smaller sections have a 50% overlap to capture segments that cross the boundaries, which are ignored by the API.

3.) Expand to neighbouring countries (eg. Austria, Switzerland, France)
This would provide a greater sample size of cities and more options for visits. I initially looked into this but was unable to find a single, easily accessible list of European cities with their population, latitude and longitude. Merging multiple lists would have increased the workload and was not possible in the time available.

4.) Strava Access token refresh
The Strava API requires an access token which expires after 6 hours, the code could be updated to automatically refresh and update prior to running this

## *6. Conclusion*

In conclusion I have been able to categorise cities in Germany based on characteristics of their eating out and running opportunities

However, the main objective of this project was to increase my knowledge of data science which I have successfully achieved. I learnt a lot about Python and associated libraries such as sci-kit learn for machine learning, matplotlib for graphs and visualisation, and folium for mapping. I also learnt how to access 3rd party data sources and wrangle data into formats required for analysis. Finally, I learned about analysing data and subsequent interpretation of the results.

## *7. References*

[1] Wikipedia list of German cities by population.
https://en.wikipedia.org/wiki/List_of_cities_in_Germany_by_population

[2] Strava API - http://developers.strava.com/

[3] Foursquare API - https://developer.foursquare.com/