# Data Science Capstone – Project Description

## Introduction

Let’s imagine we are working at a tourist agency in Germany that offers services to foreign visitors. International tourists often ask themselves which cities in their destination country they should visit. Given their limited time and resources, we need to advise them on which cities are the most worthy ones of their attention.

So far, our intuitions and expert opinions have guided the advice we give. However, data-driven and evidence-based reasoning is gaining momentum among our target audiences. They are used to online websites showing them quantified information and rankings of cities, attractions, restaurants and many other things. We can see that we are losing ground to these competitors.

So we decide to set up our own data-driven model to advise our customers on which cities in Germany they should visit. In the long run, we want to implement a sophisticated system that can be used in various ways to support our work. But for now, we start by setting up a simple analysis that we can then refine over time.

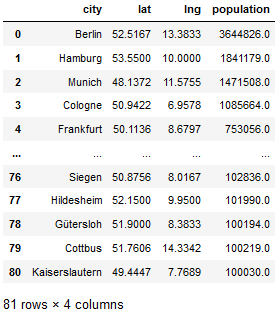
## Data

We use a dataset of major German cities, i.e. cities with more than 100,000 inhabitants. Important datapoints per city that we use for our analysis are the geographic coordinates and the population size. For that, we download a comprehensive list of all cities worldwide from the following website: <https://simplemaps.com/data/world-cities>.

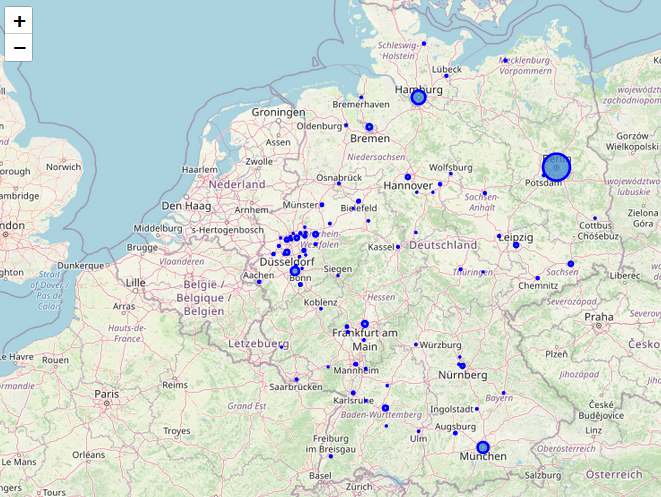
We then leverage Foursquare location data to get a list of top recommended venues around the center of each city and see how many of these venues are highly relevant to tourists (e.g. monuments, historic sites). This allows us to calculate a “Tourism Score” that indicates tourist attractiveness for each city.

## Methodology

In our first step, we rely heavily on the pandas library. We import the CSV file of all cities worldwide into a dataframe and filter for German cities with more than 100,000 inhabitants.



We can then visualize these cities on a map of Germany, using the folium library. Cities are highlighted based on their population size.

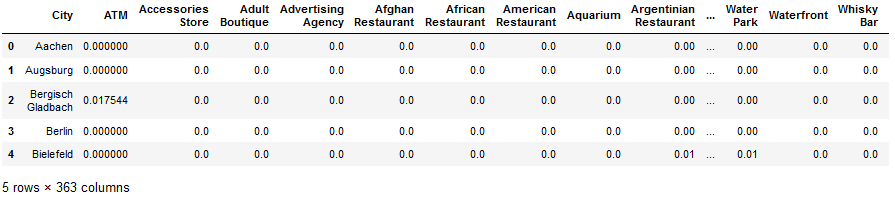


Now we can use the Foursquare API to explore these cities. We want to find the top 100 recommended venues for each city within a radius of 3 km from the city center. After cleaning the JSON file returned by Foursquare and merging it with our initial cities dataframe, we get a new dataframe containing all recommended venues. Here you can see the first five entries.

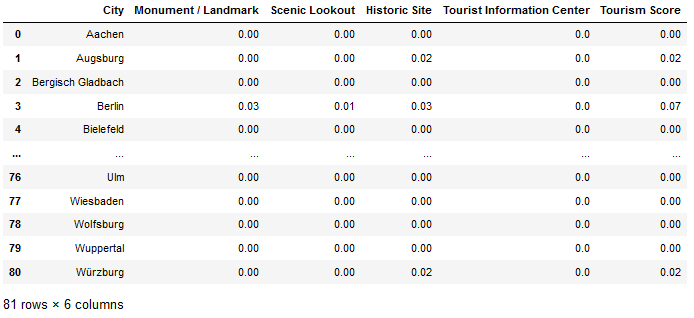


As we finally want to know which cities are most attractive to tourists, we need to define the venue categories that are relevant for this analysis. So we retrieve a list of all unique venue categories returned by Foursquare and settle on the ones most related to tourism. Four categories seem especially relevant: 'Monument / Landmark', 'Scenic Lookout', 'Historic Site' and 'Tourist Information Center'.

One option to score tourist attractiveness is to calculate the propensity of these four categories in each city. We can do this by measuring the frequency of each venue category in each city. Using one-hot encoding and grouping the dataframe by city, returns the following dataframe with all 362 unique venue categories and their frequency per city (first five rows displayed).

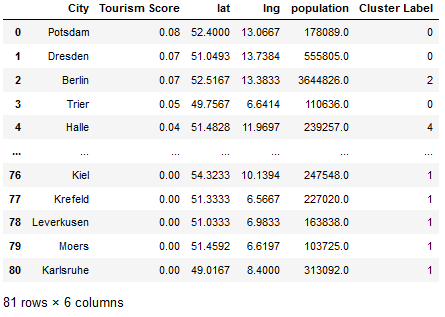


We can now create a "Tourism Score" for each city, based on the summed frequency of our four selected tourism-relevant categories.



This can also be visualized by using the matplotlib library (see results section).

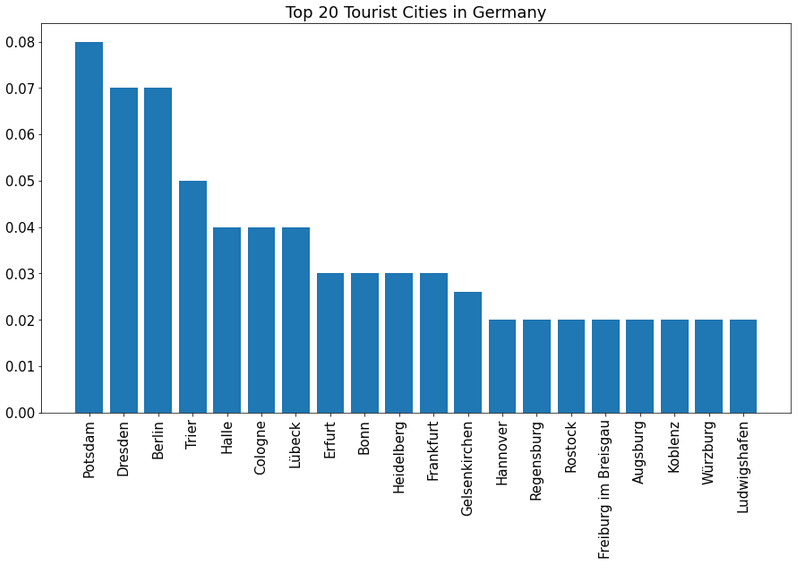
Additionally, it might be instructive to cluster cities in terms of their tourism attractiveness and population to see possible similarities between cities. Applying the k-means algorithm to create five clusters results in the following dataframe.



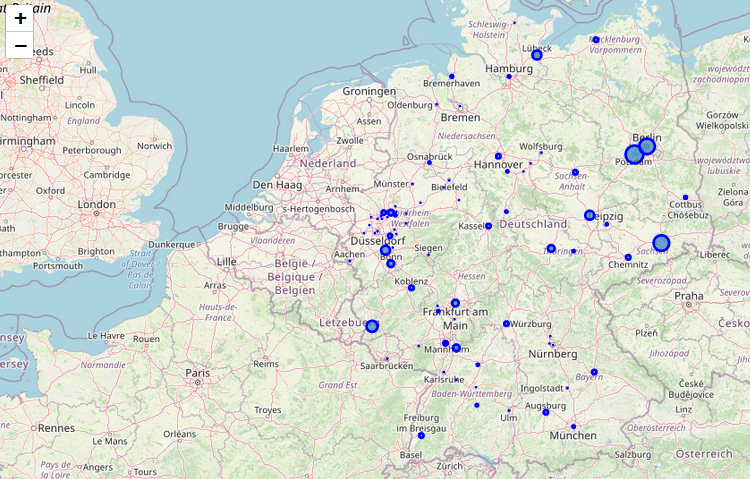
This table is of course not yet very meaningful. So let’s look at the results of our whole analysis in more detail and visualize them in a more appealing way.

## Results

First, it is now possible to identify the top 20 tourist cities in Germany based on our analysis:

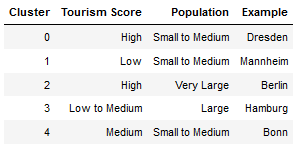


Second, we can also visualize all major German cities on a map, with the size of their marker indicating their touristic attractiveness:



We can see that there are few major cities in the East of Germany. But out of those, many seem to hold a lot of promise for tourism. Especially Berlin and its neighbor city Potsdam are rich in tourist attractions, but also Dresden and Halle score high in our analysis.

Third, we can describe the clusters returned by the k-means clustering algorithm:



It is interesting to note that Berlin forms a “one-city-cluster”, as it is far from the remaining cities in the dataframe (very large population and high tourist attractiveness). Cluster 0 and 3 are also very small, containing only three cities respectively. While cluster 0 includes the tourism-heavy cities Dresden, Potsdam and Trier, cluster 3 includes the largest German cities after Berlin, namely Cologne, Hamburg and Munich. The other two clusters are much larger in size and contain a wider variety of cities.

## Discussion

The analysis performed yielded a couple of interesting results. It seems interesting that some of the biggest German cities like Hamburg or Munich scored so low on our tourism attractiveness score. This might imply that we should direct our clients to smaller cities like Trier or Potsdam, which they are more likely to miss out on if they base their decision on the mere size and prominence of a city. The analysis also suggests that the East of Germany should be very attractive to tourists and should thus receive more of our attention. Still, the maps displayed above also show the enormous density of cities in the West of Germany (esp. North Rhine-Westphalia). This region should thus also merit a visit, even if only few of the cities located there (like Cologne or Bonn) score well in terms of tourism attractiveness.

However, all of this comes with a caveat. We need to acknowledge that this analysis only considered one variable to assess tourist attractiveness: namely the frequency of tourism-related venues returned by Foursquare. Further analysis would be needed in order to make the analysis more robust. We could for instance leverage TripAdvisor data and other sources to get a more accurate picture of tourist attractiveness.

## Conclusion

Our project shows how even a very simple analysis based on a small set of data sources can yield very interesting results. From here, we could further refine our analysis and also introduce related but independent products. For instance, we could consider setting up a personalized city-recommender system that takes the client’s preferences into account in the scoring. Such analyses were outside the scope of this particular project, but might be very valuable to work on in the future. We can the almost limitless possibilities that a data-science-driven approach can provide to our fictional business.