Coursera Capstone - Battle of the Cities (Week 1)

1. Description of the Problem and Target Audience

1.1 Description of the Problem

For multinational firms operating in different cities across the globe, it is vital to pay attention to the local cultural differences as well as tastes. For example, while the sold products might be the same, for some regions/cities recipes need to be adjusted to cater to the specific target group of that region. Moreover, marketing strategies usually depend on the people living in a city - the demographic structure as well as the 'mindset' of the region. Since granular data on a city level about the demographic structure and 'mindset' are usually scarce, an idea for companies operating in different global cities and seeking to expand to other cities would be to try to learn from the experience they already gathered.

But, how do we determine the the similarity in people and 'mindset' between cities? In their study 'The Geography of Taste: Using Yelp to study Urban Culture' (2018), Rahimi and Mottahedi showed that food choice, drink choice, and restaurant ambience can be good indicators of socioeconomic status of the ambient population in different cities. Therefore, in this paper we would like to use a city's characteristics, such as most common kinds of restaurants, amounts of parks, etc., to study the similarity of cities for companies to leverage their own experience in other cities.

1.2 Description of the Target Audience

Therefore, this paper is targeted to firms that would like to expand to new cities and leverage their experience from other regions. For example, if a firm that is currently selling products in New York and London seeks to expand to Berlin, it would be helpful for them to know whether Berlin is more similar to New York or London. Let's say Berlin is more similar to New York than it is to London. Then, the firm could adopt the same recipes and marketing strategies they already tested in New York and apply them in Berlin without having to collect massive amount of data before.

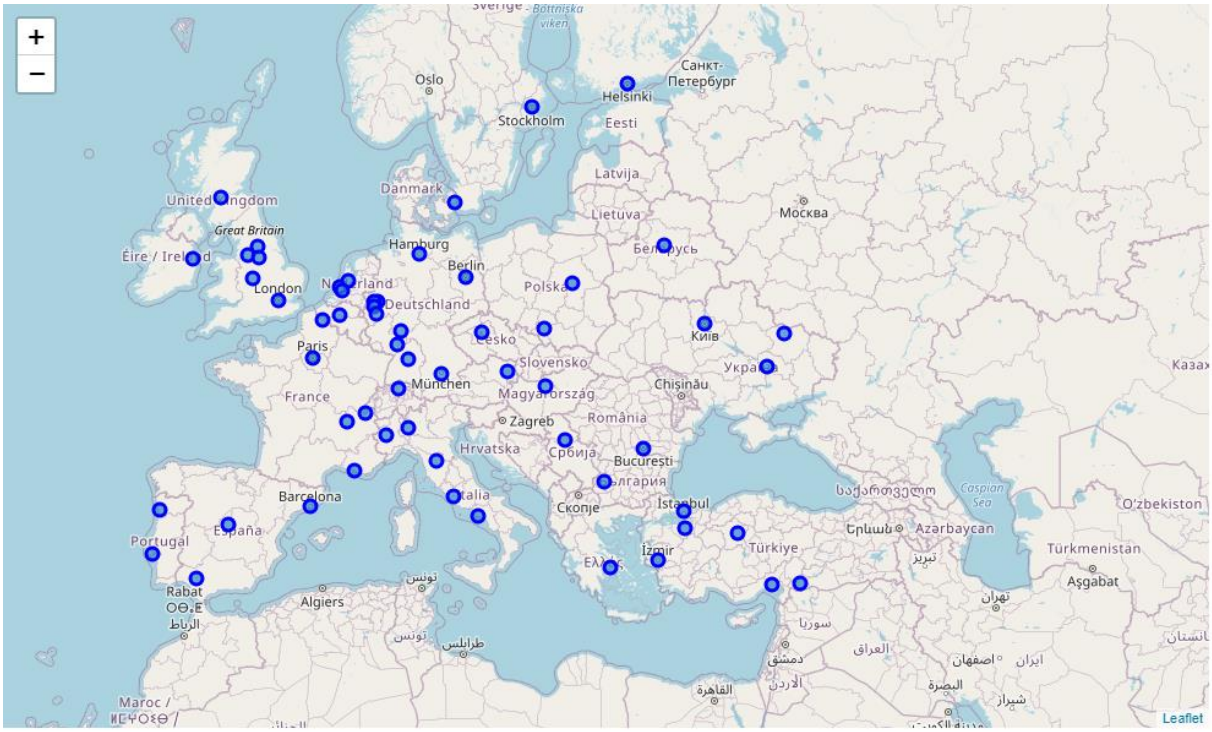
2. Description of the Data and Methodology

2.1 Data

In terms of data, I will combine two datasets:

World City Database

To get the population and geographic coordinates of the cities I take advantage of the *World City Database* that gets its data from authoritative sources such as the NGIA, NASA and the US Census Bureau: https://simplemaps.com/data/world-cities. I will focus on the biggest cities in Europe - i.e. cities with more than 1 million inhabitants. In total, I will compare these 59 big European cities:



Foursquare API

For the cities' characteristics, I will use the *Foursquare API*. With Foursquare you can download the most common sites/venues for each city. These venues can be categorized by their type, such as restaurants, parks or museums.

2.2 Methodology

In order to determine the similarity between cities I will rely on common clustering methods, such as *k-means*. I will base the clustering method on the most common sites and restaurants in each city obtained from *Foursquare*. Using the *One-hot-encoding* I will turn the venue data into a frequency with which these venues appear in each city. I will then base the clustering on the ten most common venues for each city to determine the similarity between them.