Coursera Capstone - Battle of the Cities

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 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. In this paper, I would like to take a similar, yet different approach. Instead of looking at granular differences or similarities between cities, such as different kinds of restaurants, I will adopt a higher-level approach. Therefore, in this paper I will not look at a specific category of venues but at different high-level categories in which venues usually fall. These higher-level categories will then be used to study the similarity of cities for companies to leverage their own experience in other cities.

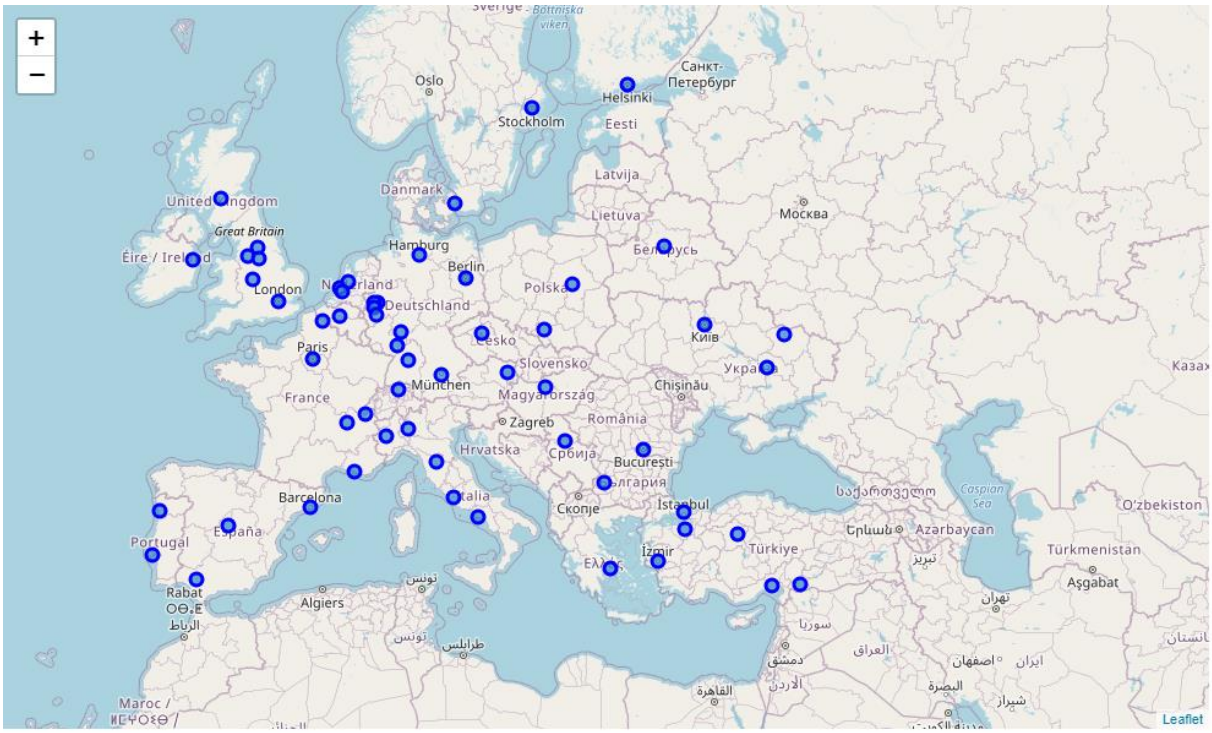
1.2 Description of the Target Audience

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

2.1 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:



2.2 Foursquare API

For the cities' characteristics, I will use the Foursquare API. Foursquare offers a list of sites/venues for each city – categorized in the following high-level categories:

Category	**ID**
Arts & Entertainment	4d4b7104d754a06370d81259
College & University	4d4b7105d754a06372d81259
Event	4d4b7105d754a06373d81259
Food	4d4b7105d754a06374d81259
Nightlife Spots	4d4b7105d754a06376d81259
Outdoors & Recreation	4d4b7105d754a06377d81259
Professional & Other Places	4d4b7105d754a06375d81259
Residence	4e67e38e036454776db1fb3a
Shop & Services	4d4b7105d754a06378d81259
Travel & Transport	4d4b7105d754a06379d81259

For each city, I will query the number of venues falling into each category in a radius of five kilometer around the city center.

3. Methodology

Before starting the actual cluster analysis, we need to query the necessary data from Foursquare for each city. For this we write a function, that loops through each city's coordinates and through each high-level category and returns the number of venues found for each category in each city in a dataframe:

	city	city Latitude	city Longitude	category	count
0	Istanbul	41.1050	29.0100	Arts & Entertainment	174
1	Istanbul	41.1050	29.0100	College & University	171
2	Istanbul	41.1050	29.0100	Event	85
3	Istanbul	41.1050	29.0100	Food	250
4	Istanbul	41.1050	29.0100	Nightlife Spot	180
5	Istanbul	41.1050	29.0100	Outdoors & Recreation	203
6	Istanbul	41.1050	29.0100	Professional & Other Places	199
7	Istanbul	41.1050	29.0100	Residence	211
8	Istanbul	41.1050	29.0100	Shop & Service	178
9	Istanbul	41.1050	29.0100	Travel & Transport	131
10	Paris	48.8667	2.3333	Arts & Entertainment	215
11	Paris	48.8667	2.3333	College & University	94
12	Paris	48.8667	2.3333	Event	11
13	Paris	48.8667	2.3333	Food	250
14	Paris	48.8667	2.3333	Nightlife Spot	241
15	Paris	48.8667	2.3333	Outdoors & Recreation	219
16	Paris	48.8667	2.3333	Professional & Other Places	179
17	Paris	48.8667	2.3333	Residence	46

Now that we have both the cities and their characteristics, we can start cleaning and preparing the data for the cluster analysis.

3.1 Data Cleaning & Explorative Analysis

Since bigger cities automatically will have more venues in almost all categories, we need to account for this fact. Therefore, for each city we calculate the relative occurrence of venues in each category by the total number of venues found.

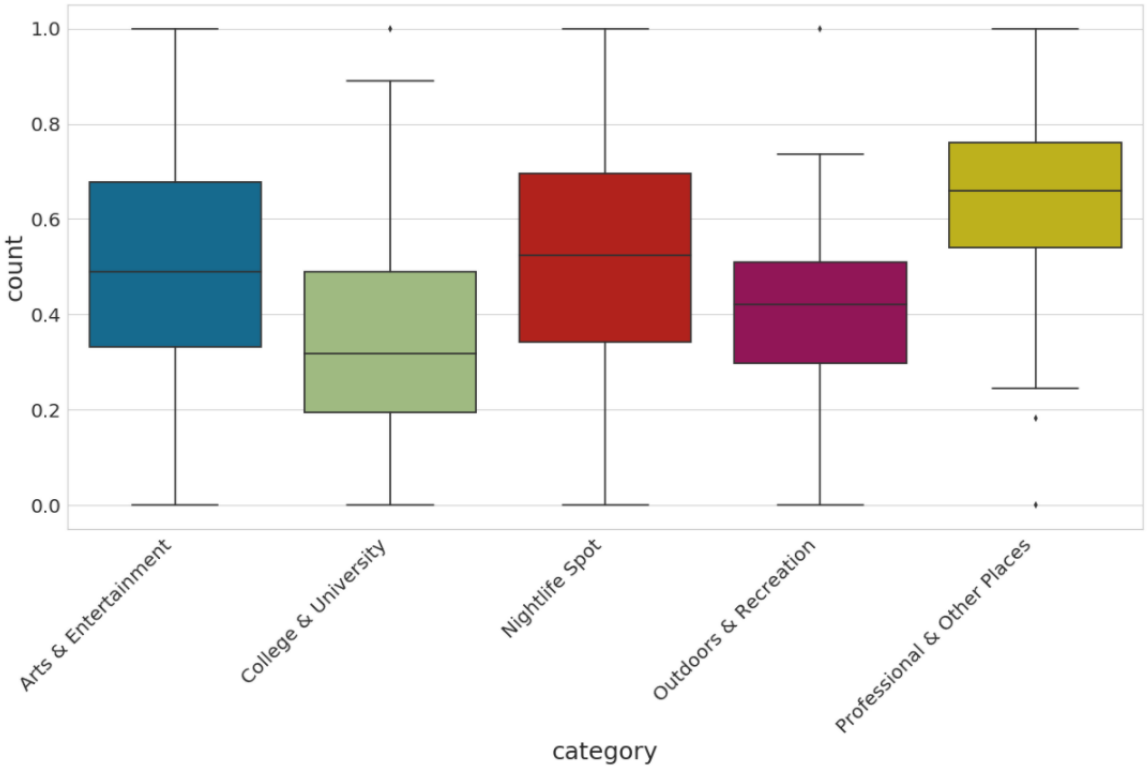
In a second step, I decide to focus on the cities in Western Europe in order to prevent too many clusters to occur. The following table shows a selection of cities and the according relative occurrences of venue categories:

category	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
city										
Amsterdam	0.110651	0.059172	0.006509	0.147929	0.144379	0.128402	0.121302	0.021893	0.127219	0.132544
Athens	0.128005	0.072716	0.005409	0.149639	0.145433	0.125601	0.127404	0.006010	0.117788	0.121995
Barcelona	0.106195	0.050374	0.006807	0.161334	0.115725	0.141593	0.140231	0.014976	0.126617	0.136147
Berlin	0.118260	0.051471	0.004289	0.152574	0.147059	0.136642	0.120711	0.014093	0.127451	0.127451
Birmingham	0.090784	0.089409	0.000000	0.125172	0.088033	0.119670	0.121045	0.034388	0.178817	0.152682
Brussels	0.099099	0.075075	0.005405	0.149550	0.149550	0.128529	0.124925	0.021021	0.122523	0.124324
Cologne	0.081699	0.062908	0.004902	0.150327	0.093137	0.149510	0.134804	0.009804	0.147876	0.165033

For the cluster analysis, we will use the variables that – in my opinion – best describe the characteristics of a city. They are

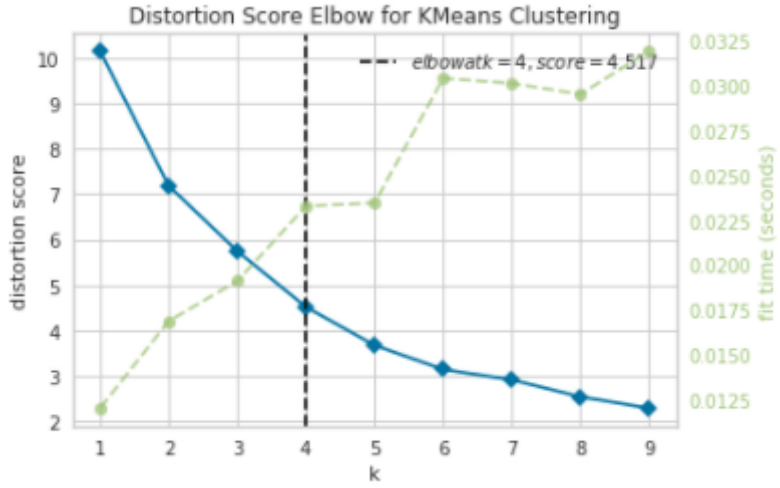
- Arts & Entertainment
- College & University
- Nightlife Spots
- Outdoors & Recreation
- Professional & Other Services

The following boxplots shows the distribution of venue categories across cities after standardizing the relative value using a *MinMaxScaler*.



3.2 Cluster Analysis

Now that we have chosen the variables along which we would like to cluster the cities, we can start the analysis. We will be using *k-mean clustering*. In order to determine the number of clusters to look for, we use the *K-Elbow-Visualizer* that plots the distortion score for different numbers of clusters.



Since there is no other clear ‘elbow’ to identify in the graph, we stick with the suggested number of *four* clusters.

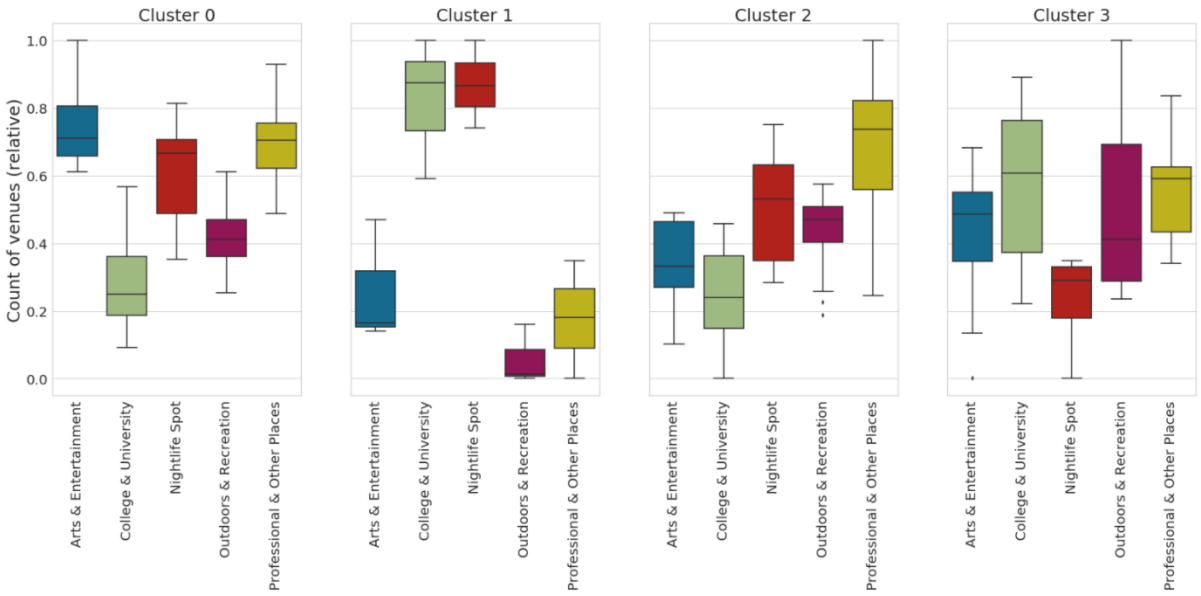
```
# set number of clusters
kclusters = 4

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(cluster_df)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

```
array([0, 0, 0, 0, 3, 0, 2, 0, 2, 3], dtype=int32)
```

4. Results & Discussion

Before we go into detail, let us look at the clustering results from afar. From a look at the map, there does not seem to be an intuitive geographical pattern along which the clusters have been formed. That makes sense, since – at least in Europe – cities are in general very similar. So how do these clusters and therefore cities differentiate?

Let us look at the different characteristics of the clusters, i.e. the relative occurrence of venue categories for each cluster:



1. **Cluster 0:** In the first cluster, we see a high level of culture, i.e. “Arts & Entertainment” as well as a high density of “Professional & Other Services” and “Nightlife Spots”. On the other hand, “Colleges & Universities” do not shape this cluster’s characteristic. Hence, these are big vibrant European cities, such as Berlin, London, Paris or Madrid, that offer great variety of culture, business and leisure.
2. **Cluster 1:** The second cluster has a high share of “Colleges & Universities” and a vibrant “Nightlife” defines the second cluster of cities. This is the smallest cluster and is geographically limited to the United Kingdom. These cities, such as Glasgow, Leeds and Sheffield seem to offer great opportunities for students and young people looking for a good night out.
3. **Cluster 2:** The third cluster focuses on the business environment since it is characterized by a high share of “Professional and Other Services” but also offers a balanced nightlife. Cities, such as Frankfurt, Milan and Dublin are known for their business networks and attract talents from all over Europe.
4. **Cluster 3:** The last cluster seems to be the most balanced set of cities with a more or less equal share in most of the categories. These cities, such as Manchester, Stuttgart and Marseille offer opportunities for both young students as well as employees.

Of course, there are a lot more characteristics along which European cities can be clustered. However, this simple approach already provides useful insights for companies that are not familiar with the European city landscape and offers some guidelines on how to develop marketing strategies for each cluster of cities.

5. Conclusion

This article used *k-means clustering* to identify clusters of (Western) European cities to provide internationally active companies with a guideline on the similarity of cities and the people living there. Companies can leverage this knowledge to cater specific marketing strategies to different sets of cities that share enough similarity for the campaign to be successful. Of course, this analysis cannot replace a fully-fledged analysis of the demographic structure of a city. However, it can help to identify the target markets where the granular data is to be collected.