Coursera Capstone Project "Battle of the Neighborhoods"

DOMINIK FREISINGER
APRIL 2020

Introduction

The topic of this project will be city tourism in Europe. Usually people tend to travel to the commonly known and popular places like Paris, London or Rome. Our goal will be to identify similar cities in Europe, based primarily on the Foursquare data. As a result a recommendation engine for travelers could be established. When a traveller likes or dislikes a city, he can find a similar city that he may also like. This way also lesser known cities could be recommended.

Data

Data sources

The data used for this project will consist of the Foursquare Places dataset, that gives information on the number of restaurants, parks, cultural venues, etc. The feature set of a given city will be based on Foursquare's venue categories:

- Arts & Entertainment
- College & University
- Event
- Food
- Nightlife Spot
- Outdoors & Recreation
- Professional & Other Places
- Shop & Service
- Travel & Transport

Therefore a city will primarily be defined by the total number of venues in the given category. Additionally climate data (average temperature per year) will be pulled from https://en.wikipedia.org/wiki/List_of_cities_by_average_temperature#Europe. Another feature that we will

add to our dataset will be the population (i.e. the size of a city), which can be obtained here: https://worldpopulationreview.com/continents/cities-in-europe/.

The final feature set will thus contain of the Foursquare categories listed above, the average temperature and the population number. What is more, we will restrict our dataset to the 53 European capitals that can be found here: https://www.nationsonline.org/oneworld/capitals_europe.htm.

Data Preprocessing

In order to be more realistic, the number of venues per category shall be evaluated per citizen. Thus, as a preprocessing step the category values are divided by the population size. Since the latter is now included indirectly in the other variables we do not use it as a feature variable on its own. Furthermore, in order to give all the features the same weight, we perform a Min-Max scaling to bring values in the range (0,1). After observing the box plots for the attributes in Fig.1, we see that we have a lot of outliers in our dataset which could lead to inaccurate clustering.

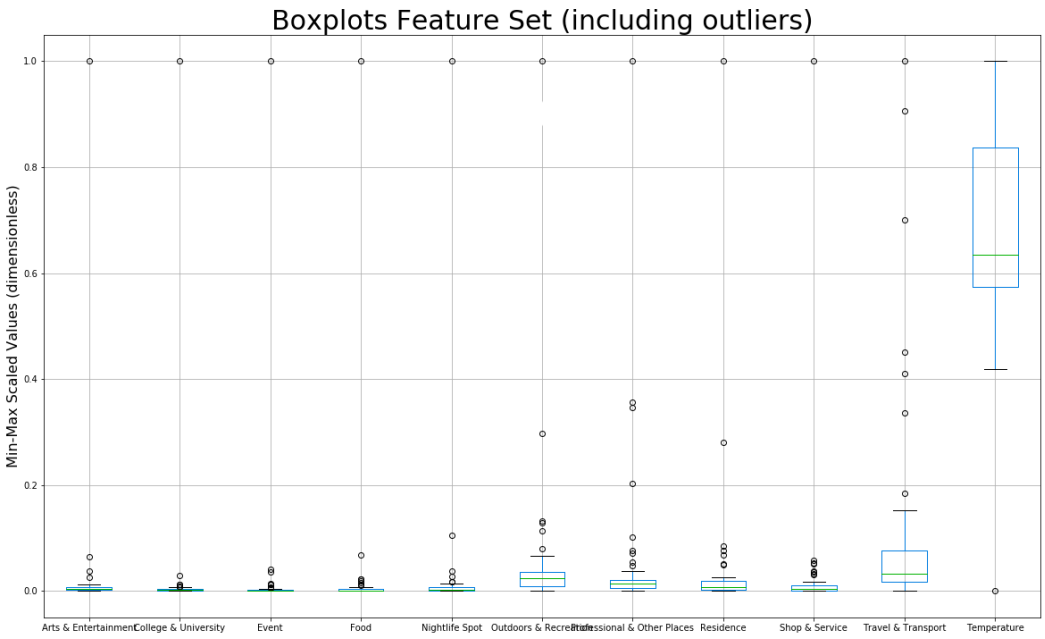

Figure 1: Boxplot of the feature variables showing a lot of outliers.

In order to reduce the number of outliers in our dataset we will drop all of the capital cities that have a population less or equal than 500 000. This reduces our dataset from 53 cities to 30. By looking at the box plots in Fig.2, we can observe that the number of outliers significantly decreased.

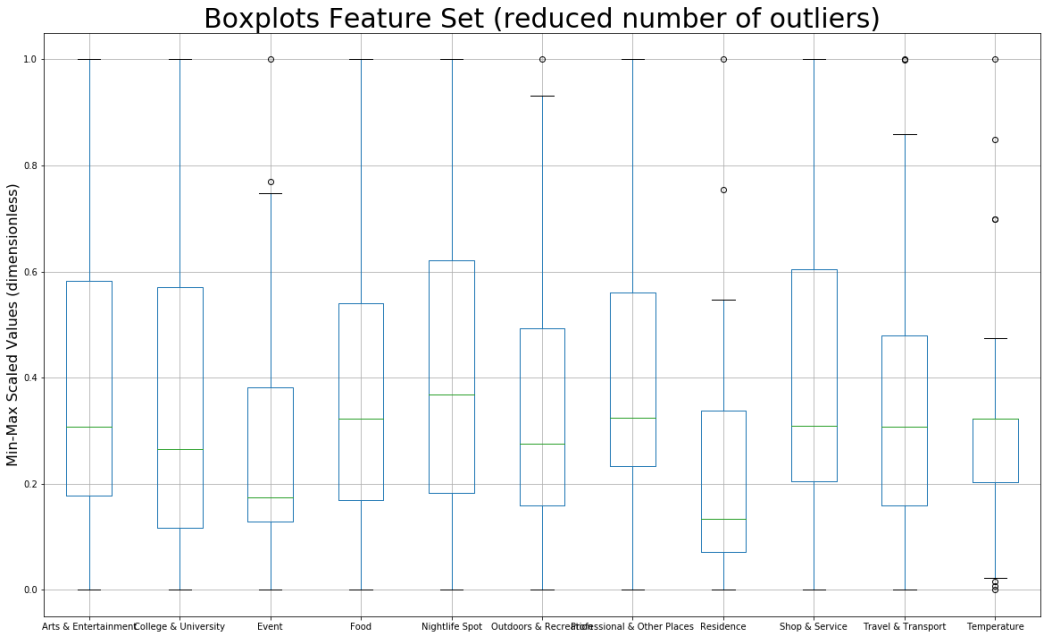

Figure 2: Boxplot after cleaning the dataset, showing less outliers.

Methodology

To find similar cities/capitals in Europe we will use a clustering algorithm. In our case, we will use KMeans and Agglomerative Hierarchical clustering and compare the outcome of those two. What is more, we will test different numbers of clusters (k) and see how the result is affected. Since we do not have any ground truth or benchmark, we cannot compare those two algorithm or say which one is better. To explore our clustered dataset we will look at the mean and variance of the different features per cluster. This can work as a measure of performance or accuracy of the clustering algorithm. We could consider the clustering algorithm (including the number of cluster) with the least standard deviation in its features the most precise one. In Fig.3 we can see the standard deviation averaged over all features and categories plotted for different values of k for KMeans and agglomerative clustering. KMeans clusters seem to have a slightly lower variance. However, as the number of cities per cluster decreases, so does the standard deviation. Therefore the optimal number of clusters can be found using the elbow method. The elbow method states that usually the k with the greatest change in the slope gives the best clustering. We want at least three clusters to have sufficient variety such that similar cities can be recommended to travellers. We conclude that the best clustering for our purposes shall thus be KMeans clustering with either 3 or 5 clusters.

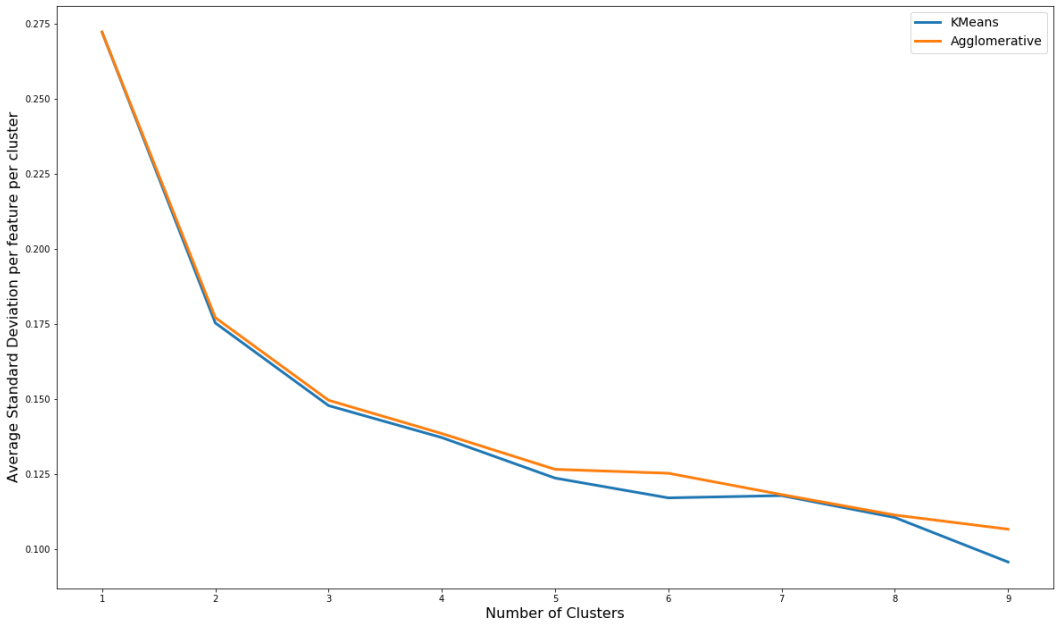

Figure 3: Average standard deviation per feature per cluster for different values of k.

Results

3-Means Clustering

In Fig. 4 we can see a map containing the different capitals that went in our cluster analysis. The color of the marker represents the category or cluster and the size is proportional the number of citizens. Tab. 1 gives an overview of the number of entities per cluster plus their average feature values. We can observe that most of the cities are in cluster 1 (marked purple in the map). Even though we did not use it as an explicit feature variable the population number is very different among the three clusters. Cluster 1 contains of the larger European capitals, like Paris, Rome, London or Moscow. In addition, these cities show a lower overall number of venues per citizen than cities in cluster 0 or 2. Cluster 0 contains cities with an average of one million citizens and a higher density of venues per citizen, such as Brussels, Zagreb or Prague. Cluster 2 comprises of the smaller European capitals, like Helsinki, Riga, Vilnius, Oslo, but also Amsterdam. We encounter the highes venue density, the top categories being “Professional and Other Places“, “Shop and Services“ and “Travel and Transport“.

Category	Count	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport	Temp	Population
0	5	0.000147	0.000128	0.000013	0.000138	0.000117	0.000227	0.000202	0.000058	0.000224	0.000194	9.500000	1.037948e+06
1	18	0.000068	0.000054	0.000004	0.000077	0.000059	0.000108	0.000097	0.000029	0.000111	0.000101	9.977778	2.607217e+06
2	7	0.000241	0.000179	0.000014	0.000222	0.000182	0.000393	0.000356	0.000085	0.000372	0.000332	9.971429	6.209827e+05

Table 1: Average feature values and total number of entities per cluster.

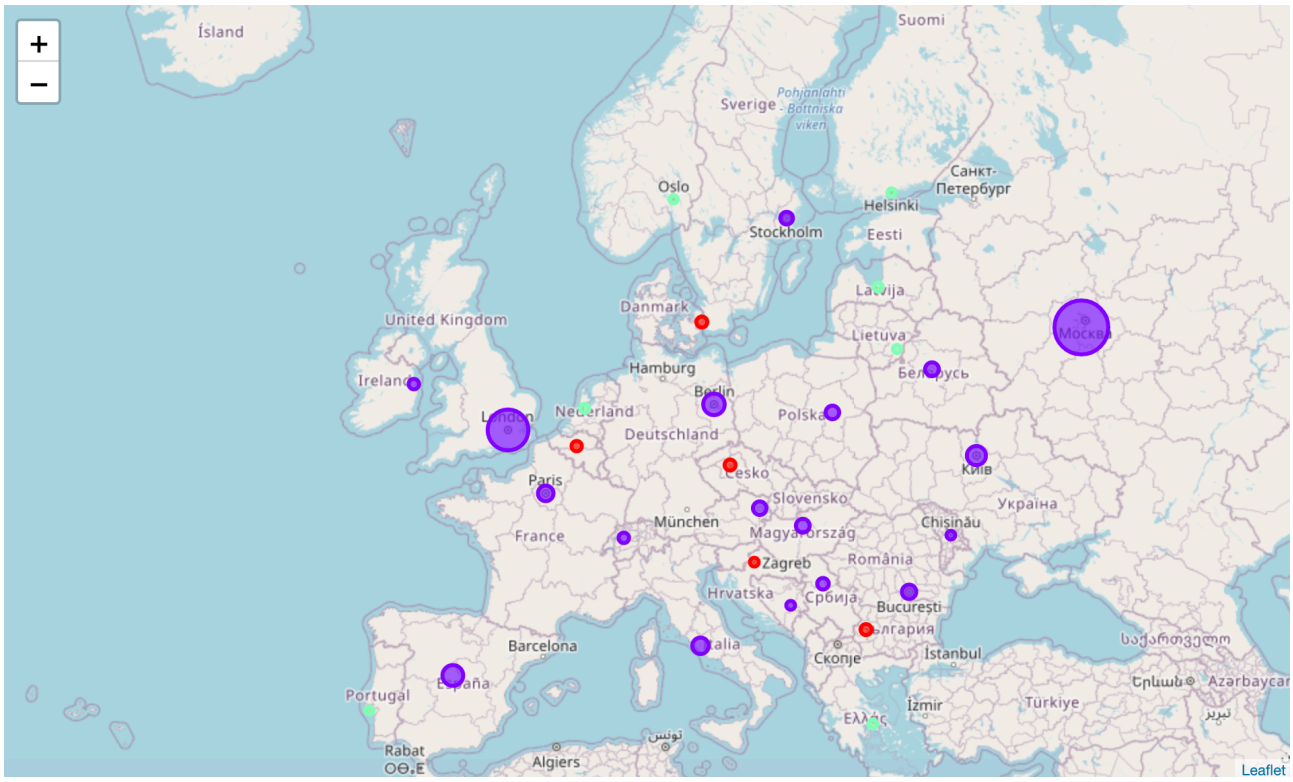

Figure 4: Map of Europe showing the capitals 3-Means clustered by color.
Red: Cluster 0, Purple: Cluster 1, Green: Cluster 2

5-Means Clustering

At a first look we can observe in Tab. 2 that the 5-Means clustering algorithm takes the temperature more into account than the 3-Means. Furthermore, we get one larger cluster (0) that includes almost 50% of the capital cities in our dataset. It consists of medium to large sized cities in terms of population and shows a moderate but not too high venue density. Cluster 1 is built from Prague, Brussels and Sofia and shows a relatively high venue density compared to cluster 0, while having similar population numbers. Cluster 2 includes the cities of Athens and Lisbon and shows the by far highest average annual temperature with 18.0 degrees Celsius. Cluster 3 consists of the largest European capitals, like London, Moscow or Paris and shows an average population count of around 4.8 million. Cluster 4 comprises the northern capitals like Oslo or Helsinki with an average annual temperature of only 6.76 degrees Celsius. In Fig. 5 we can see the five different clusters shown on a map of Europe.

Category	Count	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport	Temp	Population
0	14	0.000084	0.000069	0.000005	0.000101	0.000079	0.000139	0.000128	0.000034	0.000149	0.000132	9.492857	1.426145e+06
1	3	0.000161	0.000154	0.000017	0.000139	0.000114	0.000236	0.000216	0.000085	0.000211	0.000201	9.466667	1.112386e+06
2	2	0.000275	0.000189	0.000023	0.000236	0.000188	0.000434	0.000388	0.000065	0.000375	0.000356	18.000000	5.909240e+05
3	6	0.000051	0.000033	0.000003	0.000040	0.000033	0.000069	0.000052	0.000014	0.000066	0.000057	10.966667	4.802745e+06
4	5	0.000227	0.000175	0.000011	0.000216	0.000179	0.000376	0.000344	0.000093	0.000371	0.000322	6.760000	6.330062e+05

Table 2: Average feature values and total number of entities per cluster.

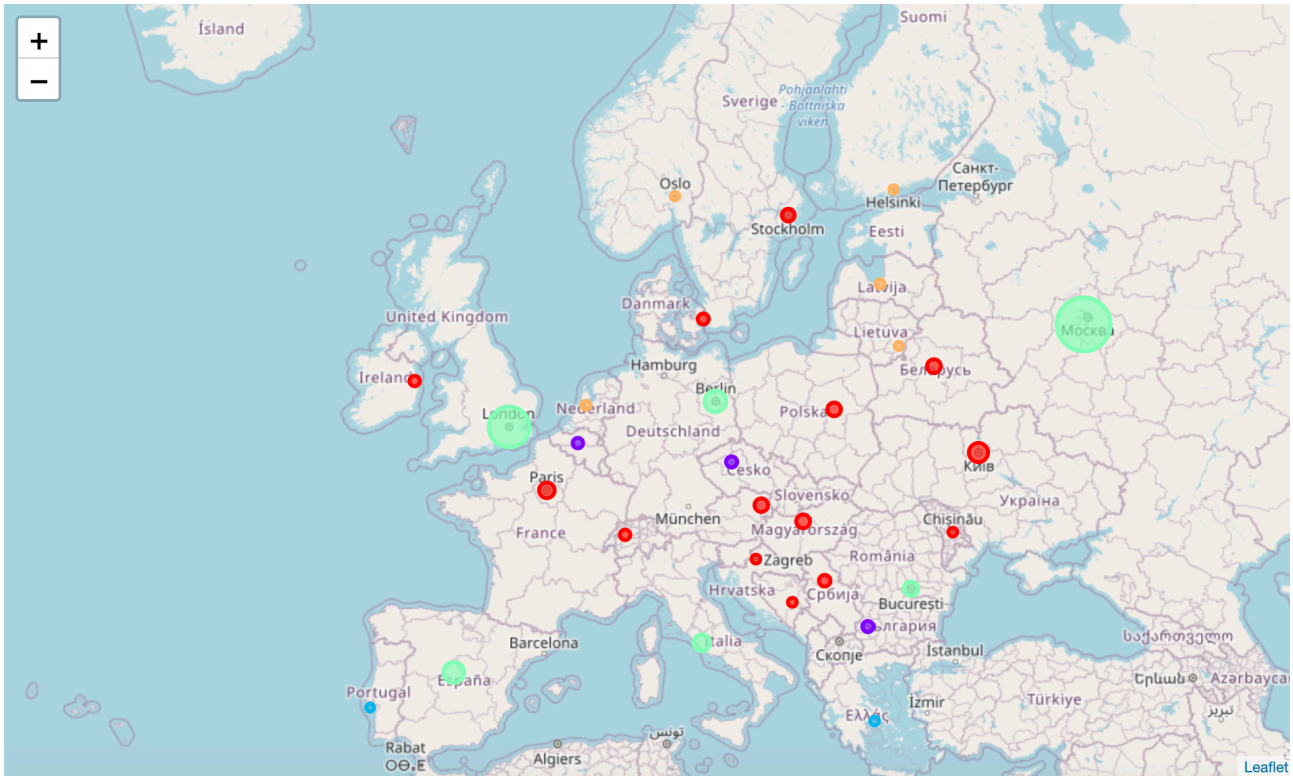

Figure 5: Map of Europe showing the capitals 5-Means clustered by color.
Red: Cluster 0, Purple: Cluster 1, Blue: Cluster 2, Green: Cluster 3, Orange: Cluster 4

Discussion

Even though the elbow method suggested that 3-Means clustering should be at least as accurate as 5-Means clustering, we see in the Results section that the latter clusters look a little more diverse. What is more, this analysis was intended to gain knowledge on similarities between European capitals such that travel recommendations can be provided. Therefore a larger number of clusters also means more diversity in our dataset and more precise recommendations. Looking at the map in Fig. 5 shows that someone who liked the city of Vienna would probably like Paris more than London. The 3-Means clustering algorithm in Fig. 4 would see Paris and London in the same cluster.

Looking at Tab. 5 we can see that the uniqueness of the different clusters is already not that bad, it should be mentioned that in order to get an even more precise result, a much larger feature set has to be taken into account. Possible parameters for future research could be the age of the city itself or the average age of its citizens because these quantities could appoint for the cities flair, which is a major criterion for travel recommendations.

Conclusion

In this project we analyzed Foursquare venue data plus parameters like population counts and average temperatures in order to cluster Europe's capital cities into different groups. The goal was to provide this clustering as a basis for possible recommendations engines for city tourism in Europe. In order to built a more precise analysis, we limited our dataset to capitals with more than 500000 inhabitants. We found that 5-Means clustering gives the best accuracy and also provides a sufficient number of clusters such that the recommendations can be more unique. The result is shown in Fig. 5. What is more, we suggested for future analysis that this feature set might has not been large enough and one also could include average age of the population or the city itself.