Capstone_report

April 6, 2021

1 Capstone Project - The Battle of the Neighborhoods

1.0.1 Applied Data Science Capstone by IBM/Coursera

1.1 Table of contents

1.2 Introduction: Business Problem

This section provides an introduction and explains the business problem that will be solved as a subpart of Coursera Capstone project.

I would like to apply clustering algorithms to biggest European cities in order to see the similarities and differences between those cities. The data will be gathered from Foursquare developer platform. Foursquare provides info about many different type of shops in the city: e.g airport shops, types of restaurants, gyms etc. A clustering algorithm can provide some insight about which cities have similar social and economic lifestyle. Such an analysis would be fun and also readers can benefit from it (e.g. decide where to live next).

The previously used Toronto report will be a very good reference for this report. I just need to decide which cities to use and their (latitude, longitude) coordinates. I will try to use cities from all around Europe. The cities should be relativeley big and popular ones. As a choosing criteria, I can use city populations or their touristic attraction. To provide more meaningful/clear results, number of cities should not be very high or very low. Around 30 city on a map can show a clear pattern. Number of clusters will be decided during implementation. The result which provides a better conclusion will be used. Based on the result and its representation on the map, more cities can be added if needed. The implementation itself will provide more insight about this.

1.3 Data

City Coordinates I decided to use city population as a metric for my city list and the list of most populated European cities can be easily reached from Wikipedia. In this section, these cities and their coordinates will be listed as dataframe and they will be shown on the map.

Also, just for curiosity, I would like to add Amsterdam as well.

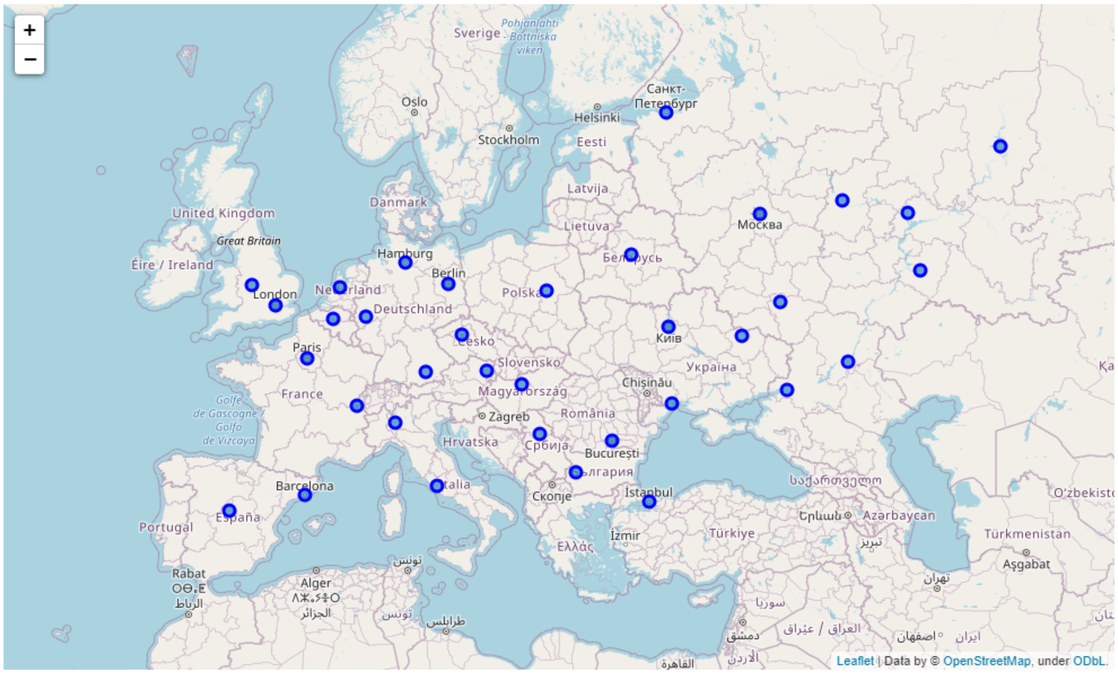
[11]:
```
df_cities
```

[11]:

	city	latitude	longitude
0	Istanbul	41.0096	28.9652
1	Moscow	55.7504	37.6175
2	London	51.5073	-0.127647
3	Saint Petersburg	59.9387	30.3162
4	Berlin	52.517	13.3889
5	Madrid	40.4167	-3.70358
6	Kyiv	50.45	30.5241
7	Rome	41.8933	12.4829
8	Paris	48.8567	2.35146
9	Minsk	53.9023	27.5619
10	Vienna	48.2084	16.3725
11	Hamburg	53.5503	10.0007
12	Bucharest	44.4361	26.1027
13	Warsaw	52.232	21.0067
14	Budapest	47.4984	19.0405
15	Barcelona	41.3829	2.17743
16	Munich	48.1371	11.5754
17	Kharkiv	49.9903	36.2304
18	Milan	45.4668	9.1905
19	Belgrade	44.8178	20.4569
20	Prague	50.0875	14.4213
21	Nizhny Novgorod	56.3286	44.0035
22	Kazan	55.7824	49.1242
23	Sofia	42.6979	23.3222
24	Birmingham	52.4797	-1.90269
25	Brussels	50.8466	4.3517
26	Samara	53.1986	50.114
27	Ufa	46.3709	6.23117
28	Rostov-on-Don	47.2214	39.7114
29	Cologne	50.9384	6.95997
30	Voronezh	51.6606	39.2006
31	Perm	58.5952	56.316
32	Volgograd	48.7082	44.5153
33	Odessa	46.4873	30.7393
34	Amsterdam	52.3728	4.8936

[13]:
```
Image("europe.png")
```

[13]:



The samples look equally distributed and suitable for clustering. Therefore, I will continue with this dataset.

Foursquare Get frequency of shop type occurence in each neighborhood. For the first 4 cities I will consider 25 km as radius of the city while for the rest I will consider 10km. This seperation is due to their big difference in population (e.g. Istanbul is 15+M while Odessa is only 1M).

Note that number of cities is one lower, because we couldnt get any information for Perm/Russia. Let's exlude it from our analysis.

1.4 Methodology

In this project I will direct my efforts on detecting clusters of big European cities which have similar characteristics. In the introduction section, I proposed to get all Foursquare data. In the Data gathering section, I decided the list of the cities and relevant Foursquare data. In this section, I realized that getting all Foursquare data for all these cities would require too many Foursquare calls, which is not allowed by my current developer subscription. Therefore, I decided to limit my focus with only Museums. The type of museums can represent alot about the culture of the city (e.g. history museums for older cities, science museum for more modern cities).

In the following Analysis section, I will apply clustering algorithms and I will try to seperate the cities with different characteristics. In the results and discussion section, I will provide the results I gathered with a good visualization. In the end, there is conclusion section in which I will summarize my findings and finish the report.

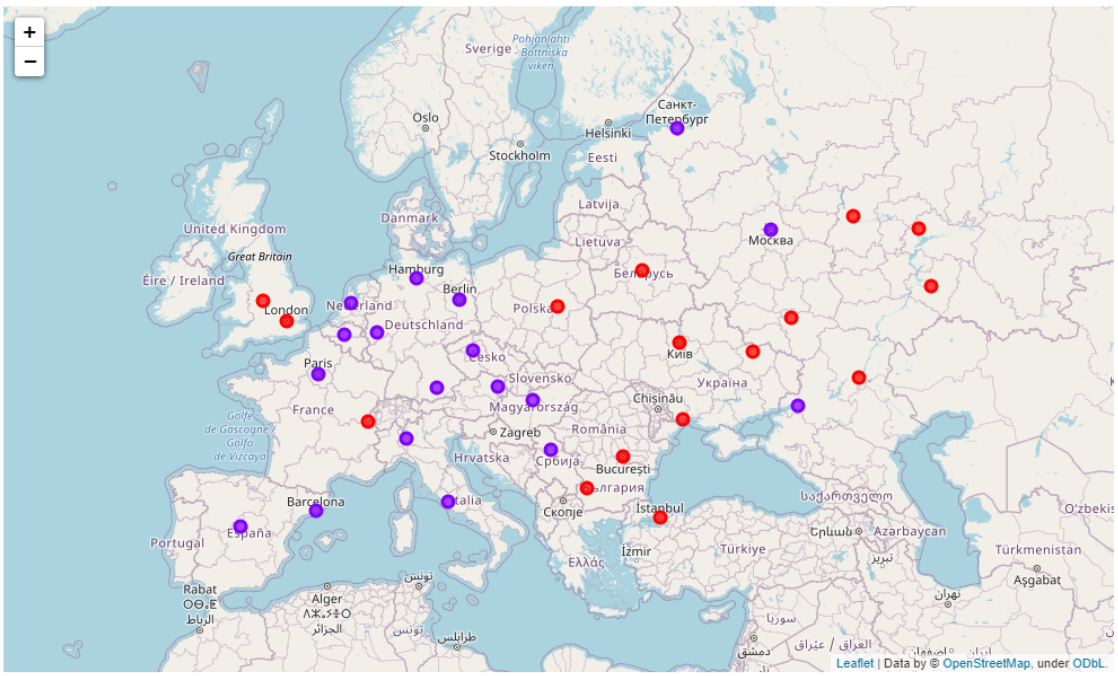
1.5 Analysis

Part of the analysis is already done in the previous sections. For example during data gathering, it was decided to limit the investigation with the Museum data. Also, the radius of city is determined considering the city population. These analysis are given in Data section because their results had effect on the data we gather. In this section, I will provide the application of the main clustering algorithm. I use k-means clustering, since it is easy to implement, fast and accurate for battle of neighborhood concepts.

Apply k-means clustering 2 clusters:

```
[28]: Image("2clusters.png")
```

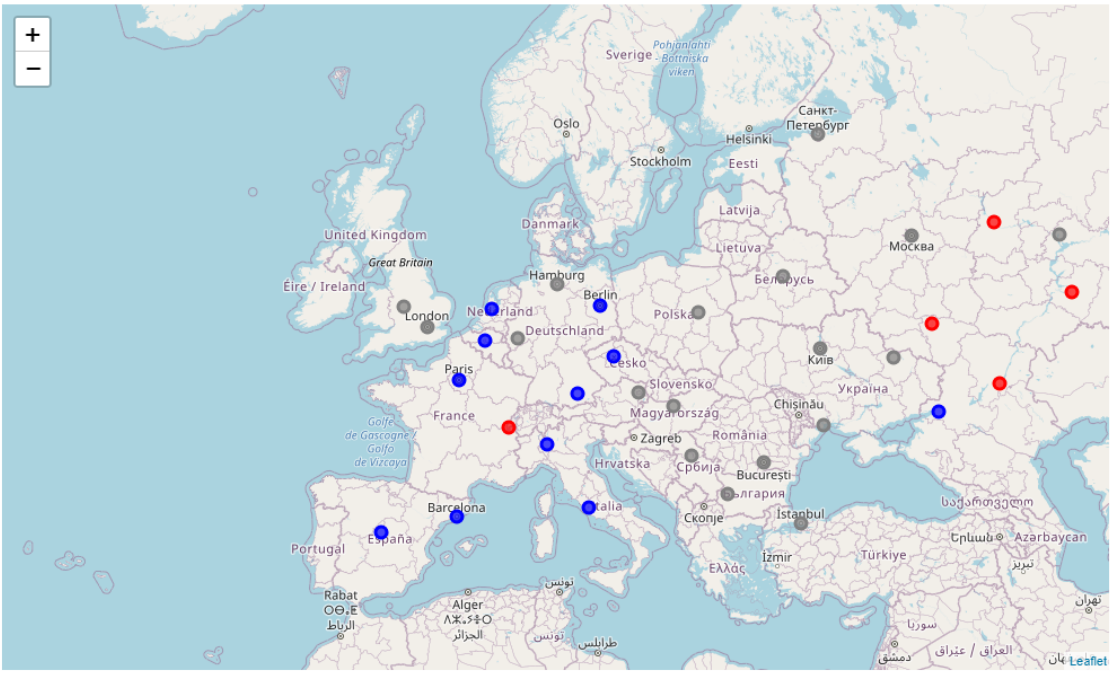
[28]:



3 clusters:

```
[39]: Image("3clusters.png")
```

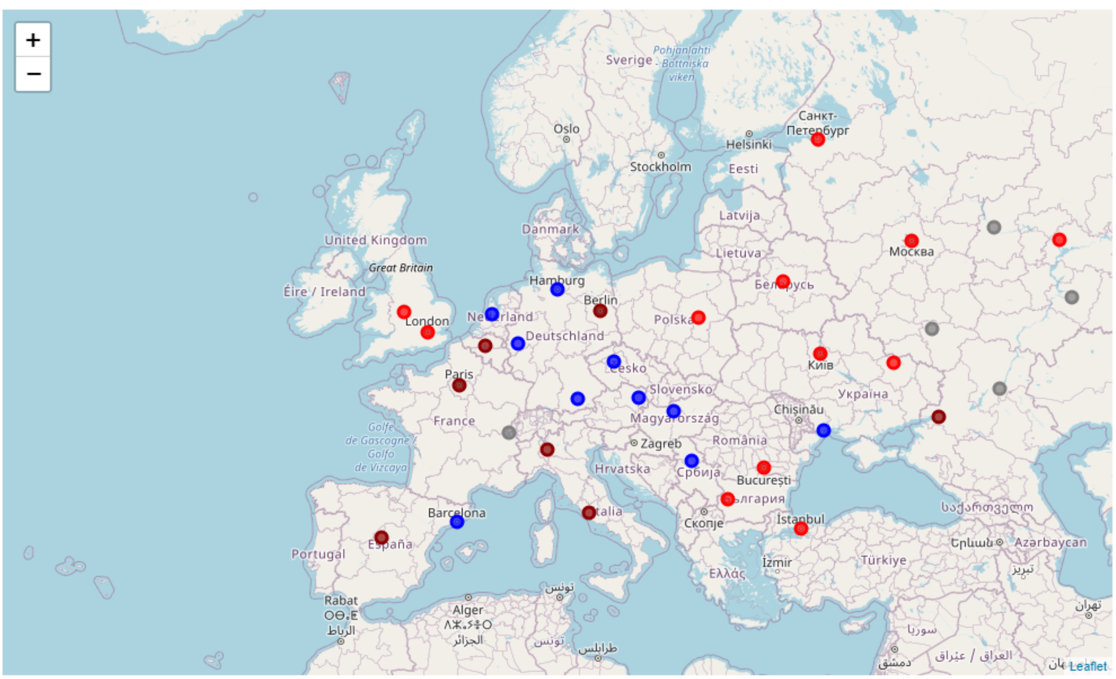
[39]:



4 clusters:

[40]:
```
Image("4clusters.png")
```

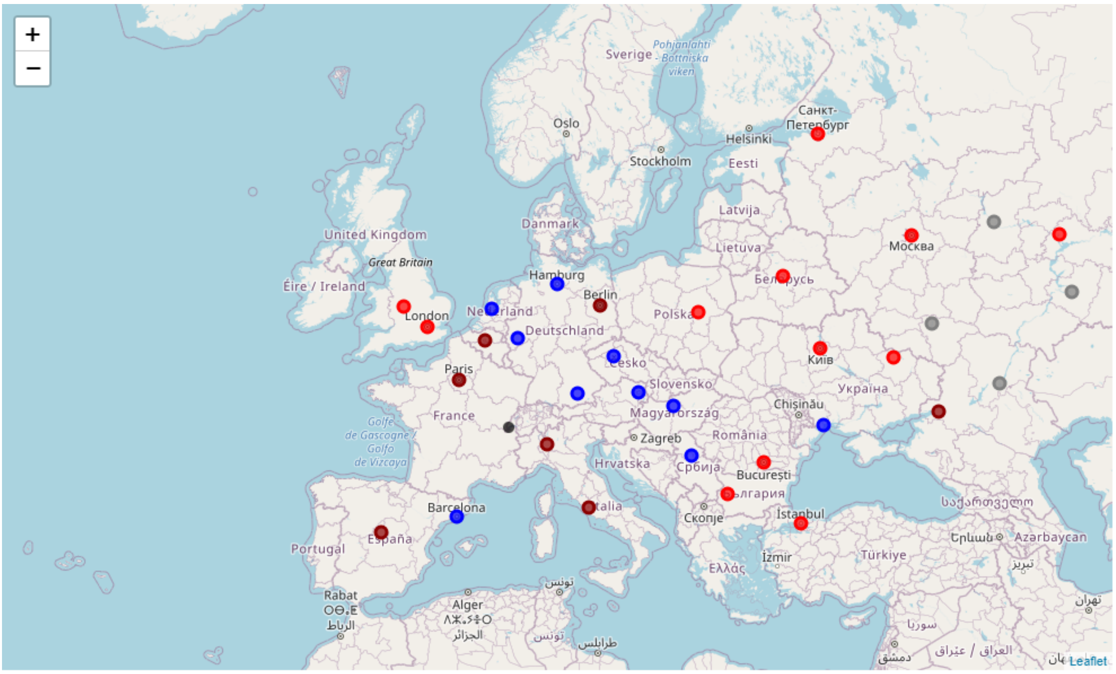
[40]:



5 clusters:

```
[41]: Image("5clusters.png")
```

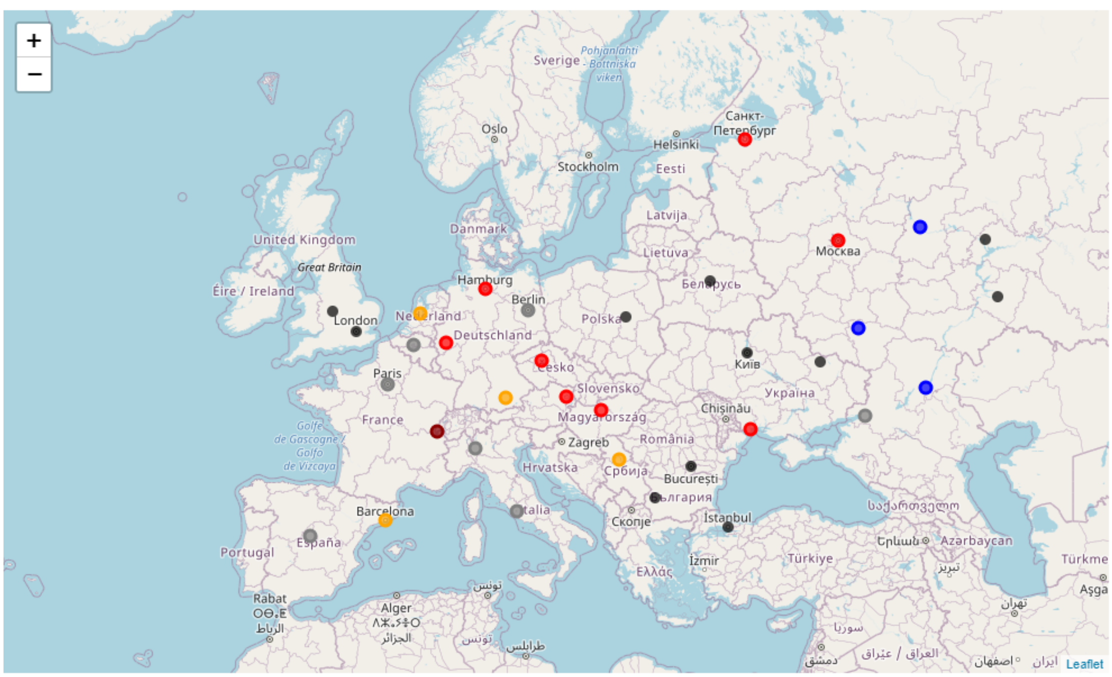
[41]:



6 clusters:

```
[42]: Image("6clusters.png")
```

[42]:



1.6 Results and Discussion

We have clusters from 2 to 6 for all European cities considering their museum data. Some interesting results are as follows:

-We have one outlier in east of France. It is always in a singled out category after 4 clusters. After some close analysis, I realized that it is actually wrong placement of city Ufa. Ufa is actually in Russia. I will ignore this sample. The reason it behaves different is because of that regions population. Since it is not a big city(due to wrong placement), it has different characteristics.

-We see that the characteristics of the museums change depending on geographic location. Usually the seperation is between east and west.

-We see that Istanbul and London shows always the same characteristics. This cannot always be explained by population, because Moscow and Saint-Petersburg have also 5M+ populations and they behave differently.

-We have same country samples for Russia, UK, Spain, Italy, Ukrain. We mostly see similar characteristics if cities belong to same country. Sometimes this does not apply for Russia because it has so many samples and big geography. Also, Barcelona and Madrid reflects some differences in Spain.

1.7 Conclusion

Purpose of this project was to identify differences between different museum types of biggest European cities. Foursquare and Wikipedia are used as data sources. Results are shown on map in Analysis section and explained in Results and Discussion section. The main result was that the museum type changes between Eastern and Western Europe. This also reflects cultural differences.

Further analysis can be also done using Restaurant types or nightlife. I believe the results I found was quite expected but still it is very much fun to verify it using data.

```
[ ]:
```