Using k-means clustering to compare cities in UK and Spain

IBM Data science Professional Certificate Capstone Report

Claire Economopoulou
May 2020

Table of Contents

Introduction ***3***

Problem **3**

Objective **3**

Target Audience **3**

Data ***4***

Methodology ***4***

Descriptive Statistics **4**

Data Wrangling **4**

Clustering **5**

Further analysis: additional requirements **5**

Results ***6***

Discussion ***7***

Conclusion ***8***

Appendix ***9***

Table of Figures

Figure 1. Clusters of cities in UK and Spain visualized on map 6

Figure 2. Cities of UK and Spain visualized on map 9

Table 1. Fist most common venues per cluster 7

Table 2. Descriptive statistics for cities, countries and data types 9

Table 3. Binary vectorization of venue categories using onehot coding 10

Table 4. Most common venues grouped by city 10

Table 5. Cities in UK and Spain grouped in clusters 10

Table 6. Cities with river, grouped in clusters 10

Introduction

The last decade has been marked as a turbulent period for the global economy which led, among others, to an increasing cross-border movement of people around the world.
International migration reports[1] show that the number of international migrants worldwide – people residing in a country other than their country of birth – reached 272 million in 2019 (from 258 million in 2017). This unpresented speed of human mobility has been largely initiated by the need of individuals for better quality of life. However, migration comes with a number of obstacles that individuals need to overcome such as social integration, closely related to cultural differences. Therefore, there is a need for migrants to move in to a place that would not be much different than their home country.

Problem

The problem that this project deals with is the difficulty for migrants to integrate in a new environment and the variety of alternatives and information sources they have before making the final decision. Under this perspective, individuals would be benefit from a way to compare cities in other countries to their own, find similarities or dissimilarities between the two and, eventually consider to move to a place that best satisfies their needs.

Objective

The objective of this project is to analyze and retrieve the cities in UK and Spain that share similar venues using data science methodology and machine learning techniques.

Target Audience

The key audience for this project is:

- individuals and families that need to move from UK to Spain and vice versa
- Companies that need to expand their businesses from UK to Spain and vice versa

[1] https://www.un.org/en/sections/issues-depth/migration/index.html

Data

The data sources that were used for this project are:

- A list of world cities, countries and their coordinates that was obtained from a World Cities Database[2] which is built using authoritative sources such as the NGIA, US Geological Survey, US Census Bureau, and NASA.
- The venues for each of the cities were obtained using the Foursquare API[3], particularly the documentation on venue categories.

Before starting with analyzing the data, the dataset was filtered to include only cities from UK and Spain and saved in csv format.

Methodology

Descriptive Statistics

A descriptive analysis of the dataset was initially performed to familiarize with the data. The dataset consisted of **94** cities, **51** for UK and **43** for Spain and a total number of **347** types of venues. The dat aset consisted of data entries of type: object (for the cities and countries), float64 (for the coordinat es – lat and lng), int64 (for the population) (Table 2, Appendix).

Data Wrangling

For the preparation of data, a set of steps was followed before proceeding to data analysis using Python:

1. The libraries needed for the data analysis were first loaded: pandas for the data analysis, numpy to handle data in a vectorized manner, sklearn.cluster for conducting k-means clustering, folium and matplotlib.cm for map visualizations.
2. The cities were first visualized on a map using folium library. Latitude and longitude coordinates of the cities found in the dataset were used in order to plot the map and also to get the venue data (Figure 2, Appendix).
3. Foursquare has one of the largest databases of 105+ million places and is used by over 125,000 developers. In order to get the venues for each of the cities, the Foursquare API was called through function that returns list of categories for venues. The venue data from Foursquare for each city, resulted in a data frame of 347 different venue categories. A limit of up to 100 venues per city and 5km radius from each location/city was applied to the function. In order to find a reasonable radius for the sample, calcmaps online tool[4] was used.
4. The function that returned the list of venues was used to assign venues to each city of the initial dataframe.
5. The venues were grouped by city and onehot encoding was performed to the new dataset so that the categorical variables are represented as binary vectors of integer values. The

[2] https://simplemaps.com/data/world-cities
[3] https://developer.foursquare.com/docs/build-with-foursquare/categories/
[4] https://www.calcmaps.com/map-radius/

mean of the frequency of occurrence of each venue category was, then, retrieved (Table 3, Appendix).

6. Finally, a new dataframe was created that displays top 10 venues for each city (Table 4, Appendix).

Clustering

In order to find which cities between the two countries are similar in terms of venues surrounding their centers, K-means clustering was performed which segmented the cities of UK and Spain in 8 clusters. A new dataframe was, then created by merging the initial dataset with the one that included the most common venues, cities, coordinates, country, population and a Cluster Labels column assigned to each city was added. The clusters were, then, visualized as colored markers on a folium map (Figure 1).

Descriptive statics on new dataframe after performing clustering analysis revealed the number of cities per cluster and, lastly, exploratory data analysis was performed, that revealed the 10 most common venues per city, all grouped by cluster(Table 5, Appendix).

Further analysis: additional requirements

In order to display only specific venues of interest per city, the new dataframe was filtered to only show those venues (examples of “River” and “Hotel” venues), keeping the cluster labelling for each of the cities that fulfil those requirements (Table 6, Appendix). Finally, a new dataframe was created and those additional requirements were isolated and visualized on a folium map with colored markers indicating the cluster the cities with the additional requirements belong to.

Results

The data analysis resulted in 8 clusters populated with cities from both countries which were visualized as a Folium map (Figure 1). Most common venues per cluster were also revealed (Table 1).

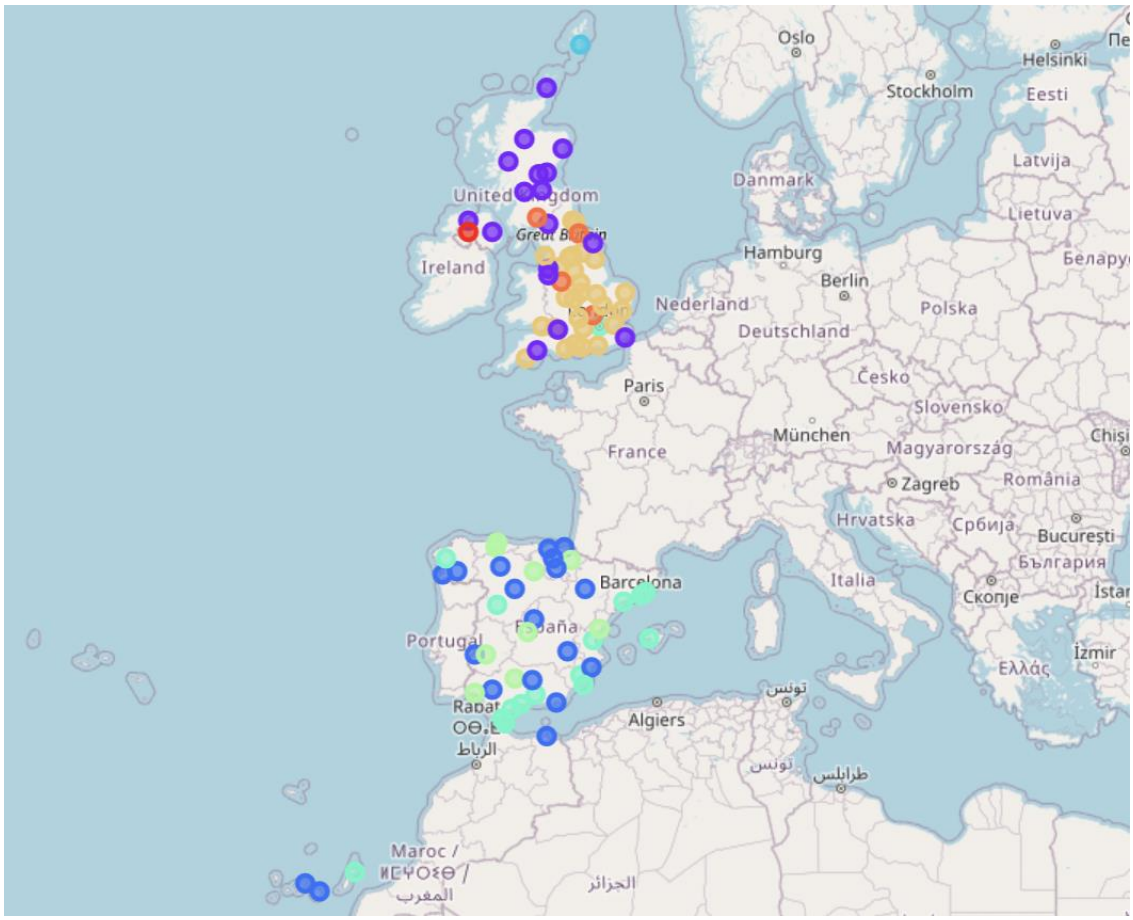

Figure 1. Clusters of cities in UK and Spain visualized on map

Clusters 0 contains 17 cities: [Glasgow, Liverpool, Edinburgh, Belfast, Aberdeen, Dundee, Exeter, Bath, Chester, Derry, Carlisle, Scarborough, Inverness, Perth, Dover, Fort William, Kirkwall] with most common venues being Pubs, Hotels, Coffe shops, Bars, Restaurants

Clusters 1 contains 19 cities: [Madrid, Sevilla, Bilbao, Zaragoza, Vigo, Las Palmas de Gran Canaria, Donostia, Santa Cruz de Tenerife, Valladolid, Alicante, Vitoria-Gasteiz, Almería, Albacete, Logroño, Ciudad de Melilla, Badajoz, León, Ourense, Jaén] with most common venues being Tapas Restaurants, Plazas, Restaurants, Hotels and Bars

Clusters 2 contains 1 city: [Lerwick] with most common venues being Harbor/Marina

Clusters 3 contains 16 cities : [London, Barcelona, Valencia, Málaga, Murcia, Granada, Palma, Cartagena, Marbella, Mataró, Salamanca, Tarragona, Algeciras, Santiago de Compostela, Ciudad de

Ceuta, Arrecife] with most common venues being Hotels, Mediterranean restaurants, Tapas Restaurants, Spanish restaurants and Plazas

Clusters 4 contains 9 cities: [Gijón, Córdoba, Pamplona, Oviedo, Castellón de la Plana, Burgos, Huelva, Toledo, Mérida] with most common venues being Spanish restaurants and Historic sites

Clusters 5 contains 27 cities : [Birmingham, Manchester, Leeds, Sheffield, Newcastle upon Tyne, Caerdydd, Nottingham, Southend-on-Sea, Bristol, Brighton, Bradford, Leicester, Sunderland, Portsmouth, Bournemouth, Coventry, Southampton, Reading, Kingston upon Hull, Blackpool, Plymouth, Oxford, Norwich, York, Ipswich, Peterborough, Cambridge] with most common venues being Pubs, Coffee Shops, bars, Indian restaurants and Cafes

Clusters 6 contains 4 cities: [Middlesbrough, Stoke-on-Trent, Luton, Dumfries] with most common venues being Pubs and Coffee shops

Clusters 7 contains 1 city: [Omagh] with most common venues being Pubs

Cluster 0		Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6		Cluster 7	
Pub	4	Tapas Restaurant	13	Harbor / Marina	1	Hotel	6	Spanish Restaurant	8	Pub	17	Pub	2	Pub	1
Hotel	4	Plaza	2			Mediterranean Restaurant	3	Historic Site	1	Coffee Shop		Coffee Shop	2		
Coffee Shop	3	Restaurant	2			Tapas Restaurant	3			Bar	2				
Bar	1	Hotel	1			Spanish Restaurant	3			Indian Restaurant	2				
Restaurant	1	Bar	1			Plaza	1			Café	1				

Table 1. Fist most common venues per cluster

For the additional requirements of locating the cities that had a river, the only city that fulfilled this requirement was Salamanca in Spain which belongs to cluster 3 (Table 5, Appendix).

Discussion

The purpose of this study was to analyze and display the cities of UK and Spain that share similarities in terms of venues. The cities were clustered using machine learning techniques and by leveraging the Foursquare API capabilities to retrieve venues for each city.

From the results, it can be assumed that UK and Spain are two countries that are dissimilar in terms of types of venues of their cities. London appears to be the only place in UK that has venues which can be also found in Spanish cities, especially those of the east coast. Additionally, although Spanish cities seem be mostly mixed regarding their types of venues, UK cities show differences between south and north UK.

A challenge that was faced during the data analysis was that a complete list of venues were not available through Foursquare API which, otherwise could yield more accuate results. Another challenge was, also, the limit in the number of venues retrieved in relation to the radius. Although Foursquare API has a limit of 100 venues per location, a number which meets requirements for a healthy data analysis, it was not deemed ideal for the current project. The reason for this was that the project aimed to analyze locations on a city-level, which requires a wider radius, something that

was not possible due to the fact that in order to produce more accurate results, the number of venues that can be retrieved should be proportionate to the radius of the city in request.

As recommendations for further study, it could be potentially valuable to consider more variables beyond venues, leveraging data related to the social landscape (e.g. sociodemographic data, data on attitudes and public opinion, crime and community safety data, social media data, data on rent prices) in order to examine urban dynamics and generate results that would better assist in decision-making.

Lastly, with variations on the data analysis approach, this project could, also, have applications in the business sector for assisting companies detect opportunities for business development in other countries, cities or neighborhoods considering nearby venues, cultural and demographic constitution of populations, trends and renting prices.

Conclusion

This study tried to cluster and compare cities in UK and Spain based on the nearby venues in order to help individuals who plan to move to a different country that suites their lifestyle and preferences. The results showed that the two countries are quite different in terms of venues in each city. There were similarities between cities of the same country but not between the cities, with London as the only exception which was clustered as unique in terms of variety of venues within the UK and similar to touristic cities of Spain's east coast. Overall, the clustering indicated that UK is quite heterogenous in relation to Spain when it comes to venue categories, resulting in the assumption that it might be easier for a person from UK to move to Spain rather than the opposite.

Appendix

```
#print the number of cities
print('Number of cities:',df.shape[0])
```

```
Number of cities: 94
```

```
df['country'].value_counts()
```

```
United Kingdom    51
Spain             43
Name: country, dtype: int64
```

```
#print the data types included in the dataframe
df.dtypes
```

```
city           object
lat           float64
lng           float64
country        object
population      int64
dtype: object
```

Table 2. Descriptive statistics for cities, countries and data types

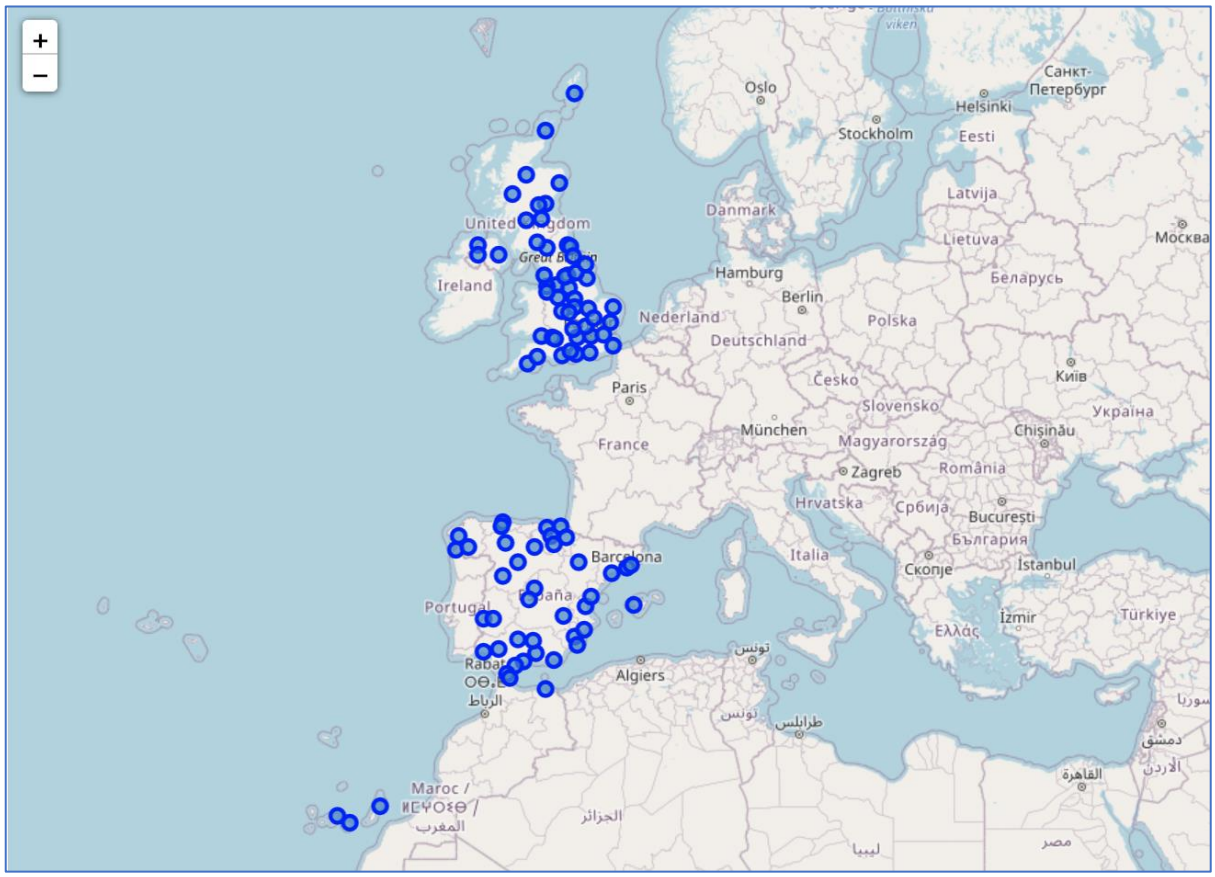

Figure 2. Cities of UK and Spain visualized on map

```
#group rows by city and by taking the mean of the frequency of occurrence of each category
groupedvenues = eu_onehot.groupby('city').mean().reset_index()
groupedvenues
```

	city	Accessories Store	African Restaurant	Airport	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Aquarium	Arcade	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Arts & Entertainment
0	Aberdeen	0.00	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1	Albacete	0.00	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
2	Algeciras	0.00	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.014706	0.000000	0.000000	0.000000	0.014706
3	Alicante	0.00	0.00	0.000000	0.000000	0.000000	0.000000	0.014493	0.00	0.000000	0.014493	0.000000	0.000000	0.000000	0.000000
4	Almería	0.00	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

Table 3. Binary vectorization of venue categories using onehot coding

[21]:

	city	lat	lng	country	population	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	London	51.5000	-0.1167	United Kingdom	8567000	3	Hotel	Theater	Cocktail Bar	Clothing Store	Grocery Store	History Museum	Park	Art Museum	Garden	Wine Bar
1	Madrid	40.4000	-3.6834	Spain	5567000	1	Plaza	Tapas Restaurant	Park	Restaurant	Art Museum	Art Gallery	Museum	Hotel	Coffee Shop	Café
2	Barcelona	41.3833	2.1834	Spain	4920000	3	Hotel	Park	Plaza	Pizza Place	Café	Dessert Shop	Bakery	Historic Site	Tapas Restaurant	Gastropub
3	Birmingham	52.4750	-1.9200	United Kingdom	2285000	5	Indian Restaurant	Pub	Coffee Shop	Bar	Hotel	Fast Food Restaurant	Restaurant	Middle Eastern Restaurant	Park	Concert Hall
4	Manchester	53.5004	-2.2480	United Kingdom	2230000	5	Coffee Shop	Pub	Park	Bar	Hotel	Beer Bar	Pizza Place	Café	Sandwich Place	Vegetarian / Vegan Restaurant

Table 4. Most common venues grouped by city

	Cluster Labels	list of cities
0	0	[Glasgow, Liverpool, Edinburgh, Belfast, Aberdeen, Dundee, Exeter, Bath, Chester, Derry, Carlisle, Scarborough, Inverness, Perth, Dover, Fort William, Kirkwall]
1	1	[Madrid, Sevilla, Bilbao, Zaragoza, Vigo, Las Palmas de Gran Canaria, Donostia, Santa Cruz de Tenerife, Valladolid, Alicante, Vitoria-Gasteiz, Almería, Albacete, Logroño, Ciudad de Melilla, Badajoz, León, Ourense, Jaén]
2	2	[Lerwick]
3	3	[London, Barcelona, Valencia, Málaga, Murcia, Granada, Palma, Cartagena, Marbella, Mataró, Salamanca, Tarragona, Algeciras, Santiago de Compostela, Ciudad de Ceuta, Arrecife]
4	4	[Gijón, Córdoba, Pamplona, Oviedo, Castellón de la Plana, Burgos, Huelva, Toledo, Mérida]
5	5	[Birmingham, Manchester, Leeds, Sheffield, Newcastle upon Tyne, Caerdydd, Nottingham, Southend-on-Sea, Bristol, Brighton, Bradford, Leicester, Sunderland, Portsmouth, Bournemouth, Coventry, Southampton, Reading, Kingston upon Hull, Blackpool, Plymouth, Oxford, Norwich, York, Ipswich, Peterborough, Cambridge]
6	6	[Middlesbrough, Stoke-on-Trent, Luton, Dumfries]
7	7	[Omagh]

Table 5. Cities in UK and Spain grouped in clusters

	Cluster Labels	city	River
76	3	Salamanca	0.013333

Table 6. Cities with river, grouped in clusters