May 2020

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

IBM Data Science Professional Certificate, Capstone project

Claire Economopoulou

Introduction

Problem

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 of choosing the right place to live in another country.

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 to help individuals find similarities or dissimilarities between their home country and the country they are moving to.

Introduction

Target Audience

- 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

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** 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**, particularly the documentation on venue categories.

Methodology

Data wrangling

1. Libraries used: pandas, sklearn.cluster, folium, matplotlib.cm.

2. The cities were first visualized on a map. 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. The categories of venues for each of the cities were retrieved using the Foursquare API (llimit of up to 100 venues per city and 5km radius from each location/city)

4. 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.

5. A new dataframe was created that displays the 10 most common venues for each city .

Methodology

Clustering

K-means clustering was performed which segmented the cities of UK and Spain in 8 clusters.

A new dataframe was created which included the list of cities, most common venues, coordinates, country, population. A Cluster Labels column assigned to each city was added.

The clusters were, then, visualized as colored markers on a folium map

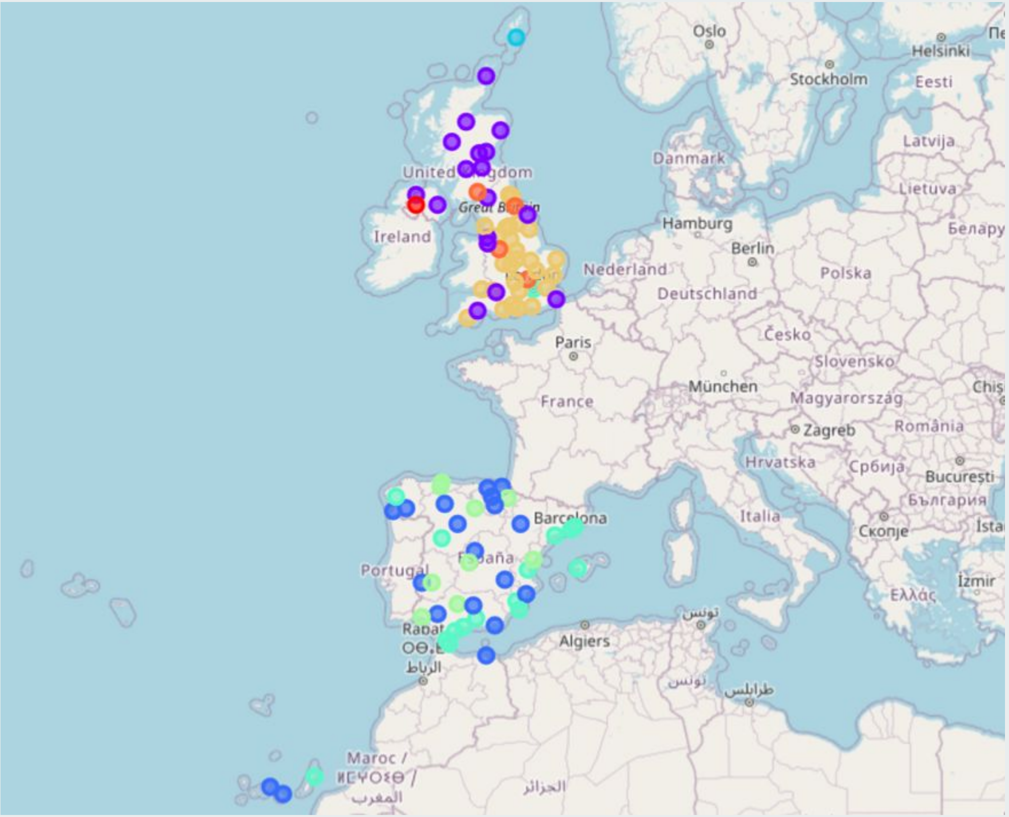
Exploratory data analysis

Exploratory data analysis revealed the 10 most common venues per city, all grouped by cluster

In order to display only specific venues of interest per city, the new dataframe was filtered and visualized in a folium map to only show those venues (e.g. “River”), keeping the cluster labelling for each of the cities that fulfil those requirements

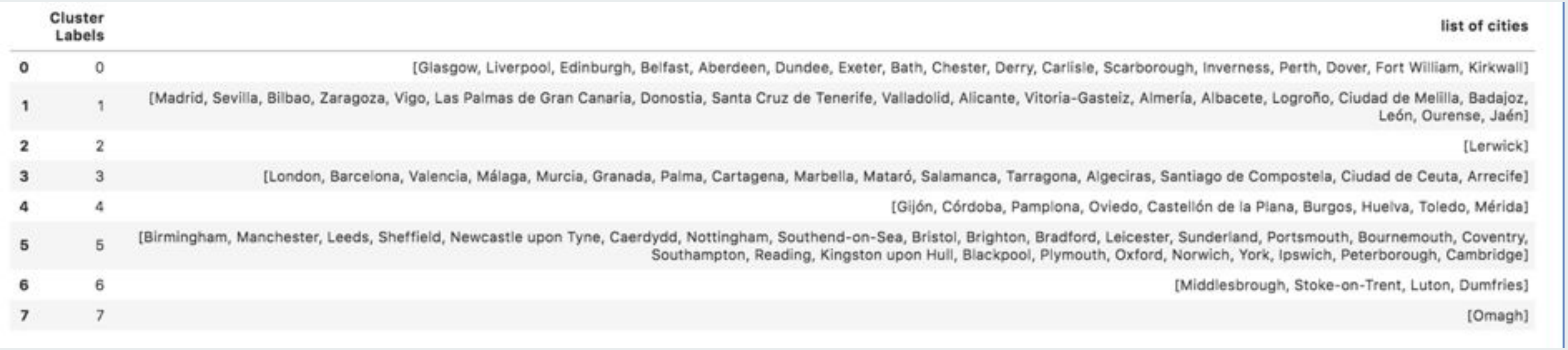
Results

The data analysis resulted in 8 clusters populated with cities from both countries which were visualized as a Folium map



Results

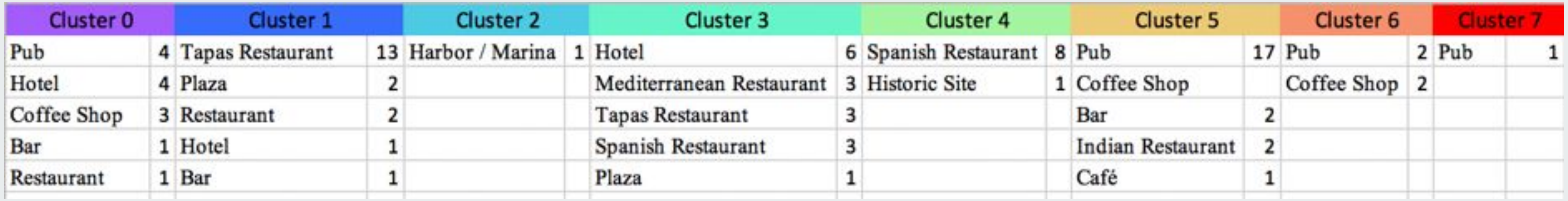
Cities in UK and Spain grouped in clusters



	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]

Results

Most common venues per cluster



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				

Discussion

Observations

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.

Challenges

- a complete list of venues were not available through Foursquare API
- the limit (100) in the number of venues retrieved in relation to the radius, numbers that should be proportionate but not ideal for the scope of the project which operates on city-wide level

Recommendations

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.

Discussion

Recommendations

- Leverage 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.

- Assist 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 suits 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.

Thank you!