2019

What is Foursquare opinion on splitting up Belgium?

CAPSTONE PROJECT FOR COURSERA IBM DATA SCIENCE SPECIALIZATION

ADRIEN ALLART

Contents

1. Introduction

An over-simplification of Belgium structure as a country could be visualize as 3 regions (Flanders in the North and Wallonia in the South, Brussels in the middle) with two different language (Dutch and French) living under one flag. The idea of splitting the country is a topic that is often discussed with a lot of arguments on both sides. It involves a lot of geo-political, socio-economic and cultural questions.

Answering whether this would or would not be a good idea is obviously out of the scope of this project as it is an extremely complex problem. The idea of this project is to compare how similar or dissimilar cities from the regions are based on location data from Foursquare. It will also extend the analysis to neighboring countries (The Netherlands and France) which are often considered as viable options for regions of Belgium with which they share a language.

This will hopefully provide another perspective and be an objective data-driven argument in the discussion. This might be interesting for decision maker that have to handle that question, might also be of use of anyone wanting to build their opinion on the question or to validate/challenge their existing opinion. It might also give some insights on how similar some of the founding members of the EU are.

More information about the partition of Belgium might be found at:
https://en.wikipedia.org/wiki/Partition_of_Belgium

2. Data

2.1. Dataset

There are two main source of data that are used in the following analysis:

The first dataset required is a list of cities for each country (Belgium – France – The Netherlands) with their longitude – latitude coordinates. The list of city with their location can be found in a Kaggle dataset from MaxMind here . The same dataset give regions for each country which for the case of Belgium matches the country provinces. That information is used to separate the country by its regions.

The raw data look like this:

```
Country,City,AccentCity,Region,Population,Latitude,Longitude
ad,aixas,Aixàs,06,,42.4833333,1.4666667
ad,aixirivali,Aixirivali,06,,42.466666700000005,1.5
ad,aixirivall,Aixirivall,06,,42.466666700000005,1.5
ad,aixirvall,Aixirvall,06,,42.466666700000005,1.5
ad,aixovall,Aixovall,06,,42.466666700000005,1.4833333
ad,andorra,Andorra,07,,42.5,1.5166667
ad,andorra la vella,Andorra la Vella,07,20430.0,42.5,1.5166667
ad,andorra-vieille,Andorra-Vieille,07,,42.5,1.5166667
ad,andorre,Andorre,07,,42.5,1.5166667
ad,andorre-la-vieille,Andorre-la-Vieille,07,,42.5,1.5166667
ad,andorre-vieille,Andorre-Vieille,07,,42.5,1.5166667
ad,ansalonga,Ansalonga,04,,42.5666667,1.5166667
ad,anyos,Anyós,05,,42.533333299999995,1.5333333
ad,arans,Arans,04,,42.5833333,1.5166667
ad,arinsal,Arinsal,04,,42.5666667,1.4833333
ad,aubinya,Aubinyà,06,,42.45,1.5
ad,auvinya,Auvinya,06,,42.45,1.5
ad,bicisarri,Biçisarri,06,,42.4833333,1.4666667
ad,bixessarri,Bixessarri,06,,42.4833333,1.4666667
ad,bixisarri,Bixisarri,06,,42.4833333,1.4666667
ad,canillo,Canillo,02,3292.0,42.5666667,1.6
ad,casas vila,Casas Vila,03,,42.533333299999995,1.5666666999999999
ad,certers,Certers,06,,42.466666700000005,1.5
ad,certes,Certés,06,,42.466666700000005,1.5
ad,eixirivall,Eixirivall,06,,42.466666700000005,1.5
ad,el pui,El Pui,04,,42.55,1.5166667
ad,els bons,Els Bons,03,,42.533333299999995,1.5833333
ad,el serrat,El Serrat,04,,42.6166667,1.55
ad,els plans,Els Plans,02,,42.5833333,1.6333333
ad,el tarter,El Tarter,02,,42.5833333,1.65
ad,el tremat,El Tremat,03,,42.55,1.5833333
```

The second set of data required for building the tool is Foursquare API that will be used to retrieve information of surrounding venues for the geolocation data chosen from the previous set of data.

The API transform longitude and latitude data into a JSON with list of venues located around the position and the category of the venue (as well as a lot of other information that are not used for this project).

Not critical for the overall result, but geojson file of Belgium provinces is used to draw the border of each provinces on a couple of map. The data was found on digitalwallonia.be.

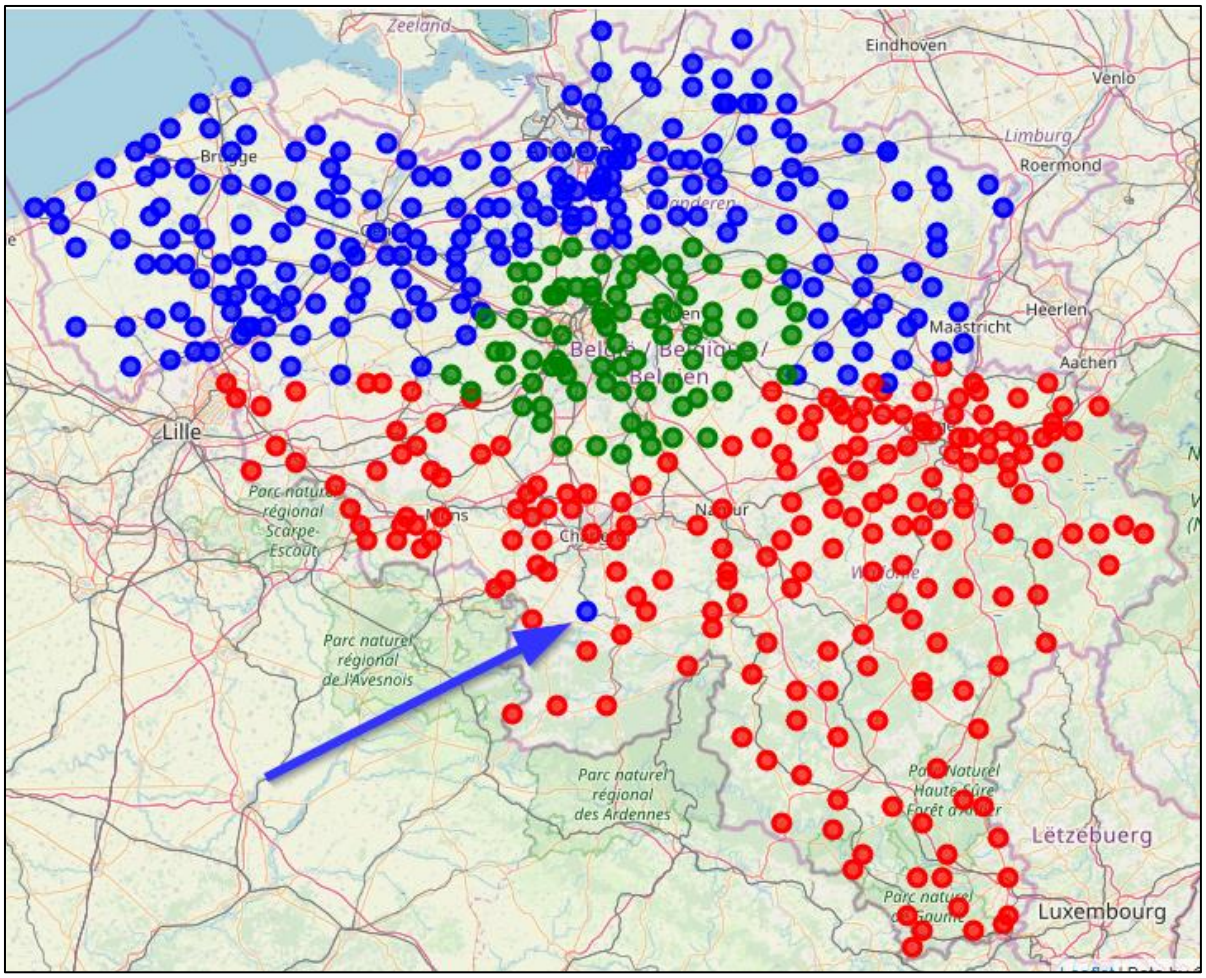
2.2. Data Cleaning / Data Manipulation

Some data cleaning has to be done on the dataset that are used:

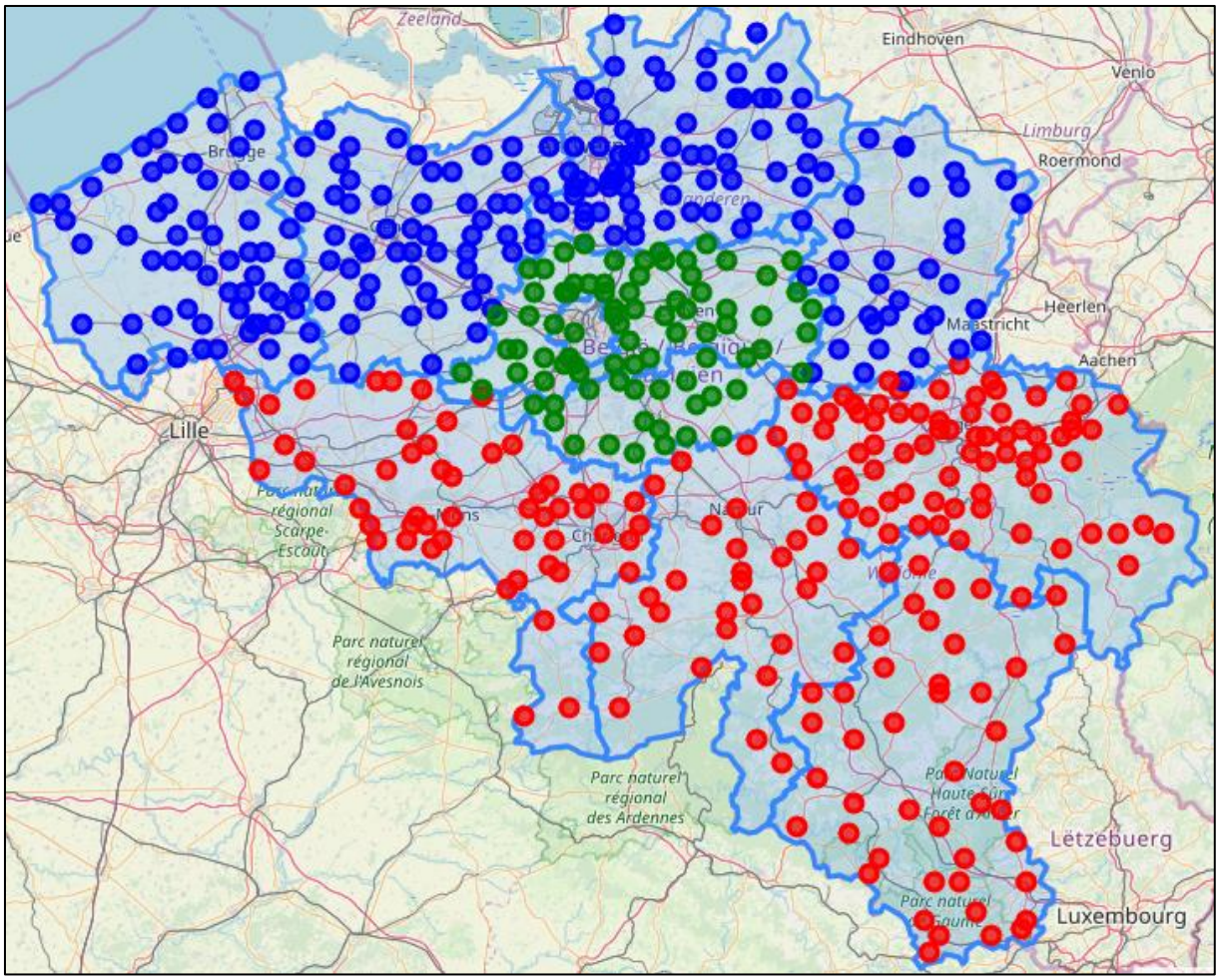
1. The cities dataset has to be rescoped to keep only the country of interest for the analysis (Belgium – France – The Netherlands)
2. The numerical value for the region corresponding to Belgium in the cities dataset had to be transform in either Flanders ,Wallonia or Brussels. There was some issue with the province Brabant which should be split into 2 (Flanders Brabant and Wallonia Brabant) but the dataset had most cities grouped into a single number with a couple cities on their own. To simplify the analysis I decided to include all of it into the Brussels region. The table I used for the transformation is as below (source: me, am Belgian):

```
01 : Antwerp                              : Flanders
02 : Brabant (except couple cities)       : BXL
03 : Hainaut                              : Wallonia
04 : Liege                                : Wallonia
05 : Limbourg                             : Flanders
06 : Luxembourg                           : Wallonia
07 : Namur                                : Wallonia
08 : Flandre or                           : Flanders
09 : Flandre occ                          : Flanders
10 : single city                          : BXL
11 : BXL                                  : BXL
12 : single city                          : BXL
```

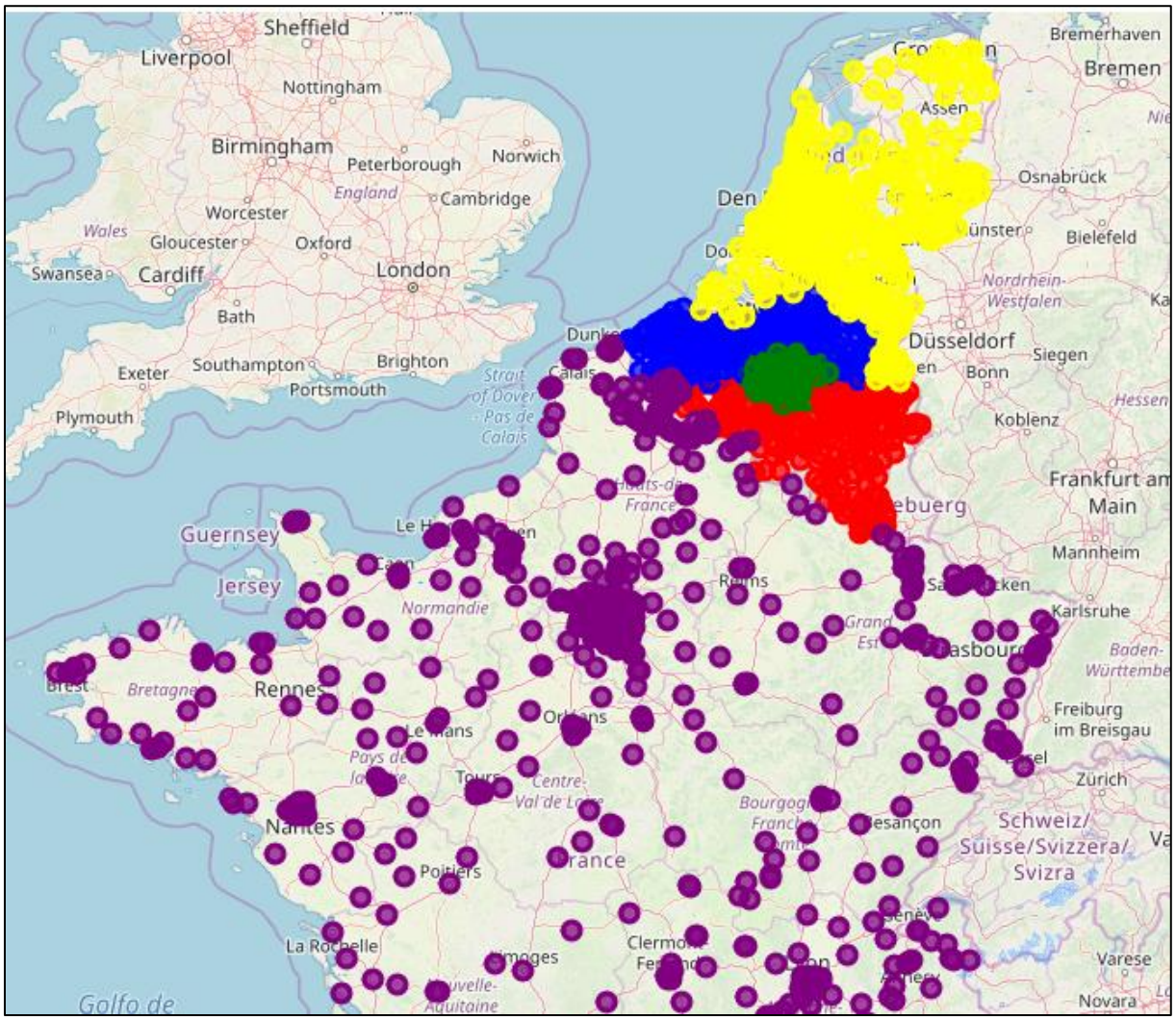
3. Visualization of the cities on the map allowed to spot an error in the dataset and fix it



Here is the data corrected with the Provinces lines:


Zeeland
Eindhoven
Venlo
Limburg
Roermond
Heerlen
Maastricht
Aachen
Lille
Parc naturel régional Scarpe-Escaut
Parc naturel régional de l'Avesnois
Parc naturel régional des Ardennes
Lëtzebuerg
Luxembourg

Similar visualization was done for the second part of the analysis with The Netherlands and France:



4. The city data is used to create a dataset of the venues using foursquare:

	City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	aalst	50.933333	4.033333	Basic-Fit	50.934496	4.034168	Gym / Fitness Center
1	aalst	50.933333	4.033333	De Prins Drinkt Koffie	50.934941	4.038068	Coffee Shop
2	aalst	50.933333	4.033333	Tang's Palace	50.935172	4.038079	Chinese Restaurant
3	aalst	50.933333	4.033333	Healthfactor Gym	50.930120	4.035189	Gym / Fitness Center
4	aalst	50.933333	4.033333	Pizza Talia	50.932481	4.030968	Pizza Place

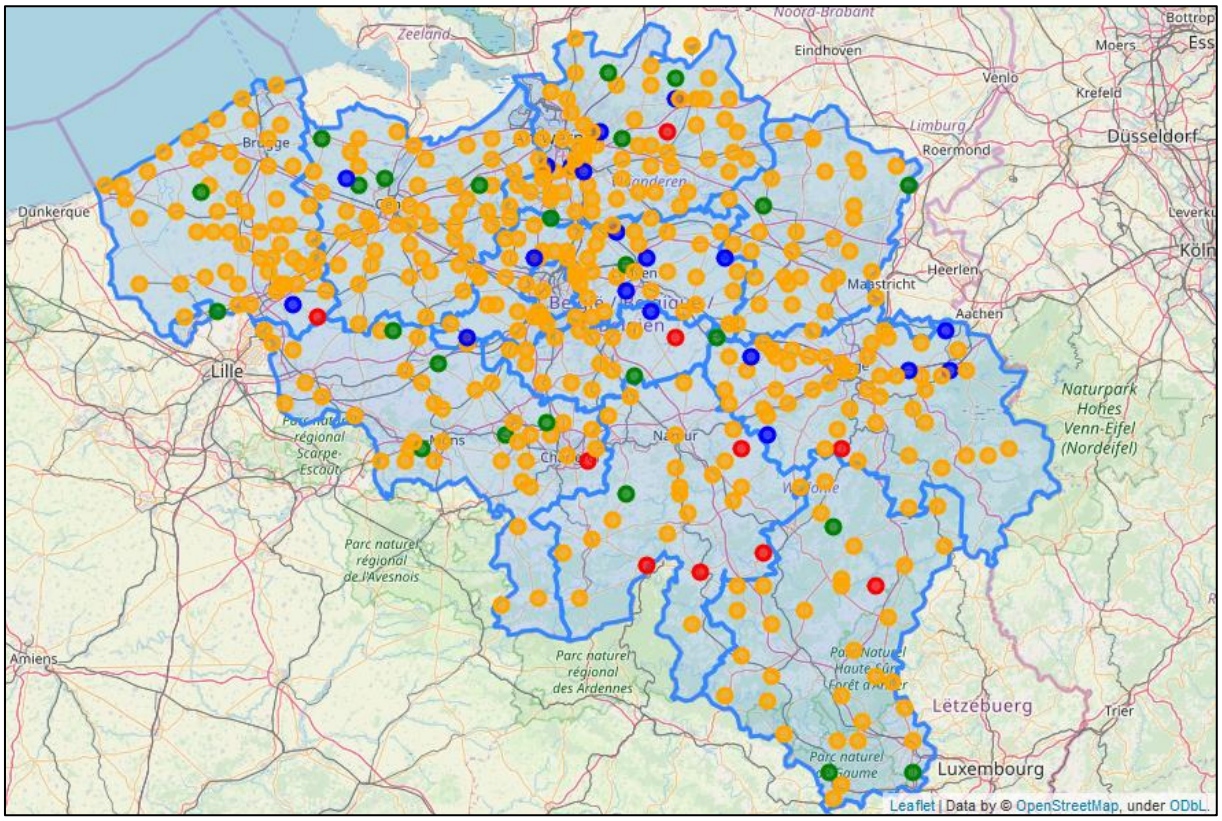
From this venue data, one hot encoding is done on the venue category column and then is group by city to display the proportion of each type of venue the city have which will be the features for the model.

3. Methodology

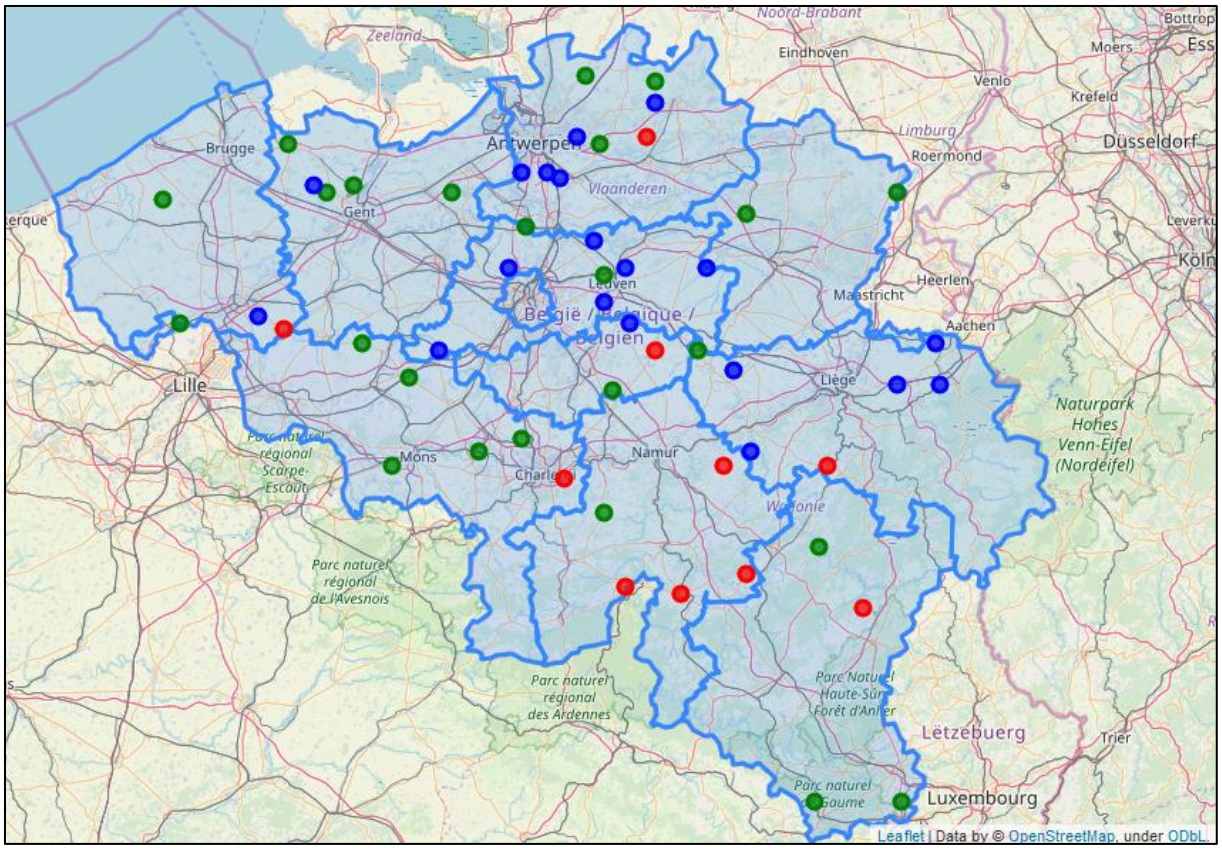
The methodology that will be adopted here will be to use K-mean clustering to try and separate all the cities in different clusters. Once this is done, an analysis of the proportion of each cluster will be done in Belgium regions at first and combine with France and The Netherlands after.

4. Results

The result of the K-Mean clustering for Belgium can first be observed as follow:



The obvious observation is that the majority of cities for the whole country seems to be from the same cluster. There is however some other clusters, let's remove the most popular one to see those better:

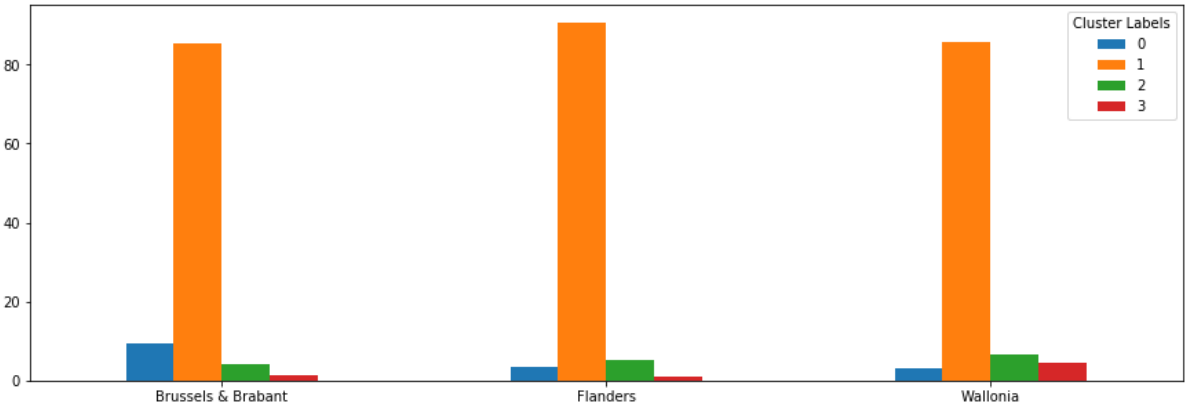


Here we can see the green and blue cluster present at a lot of places in the country while the red cluster seems more present in Wallonia.

Putting the different proportion of clusters for each region give the following table:

NewRegion Cluster Labels	Brussels & Brabant	Flanders	Wallonia
0	9.33	3.27	3.21
1	85.33	90.65	85.90
2	4.00	5.14	6.41
3	1.33	0.93	4.49

Which put into graphs gives:

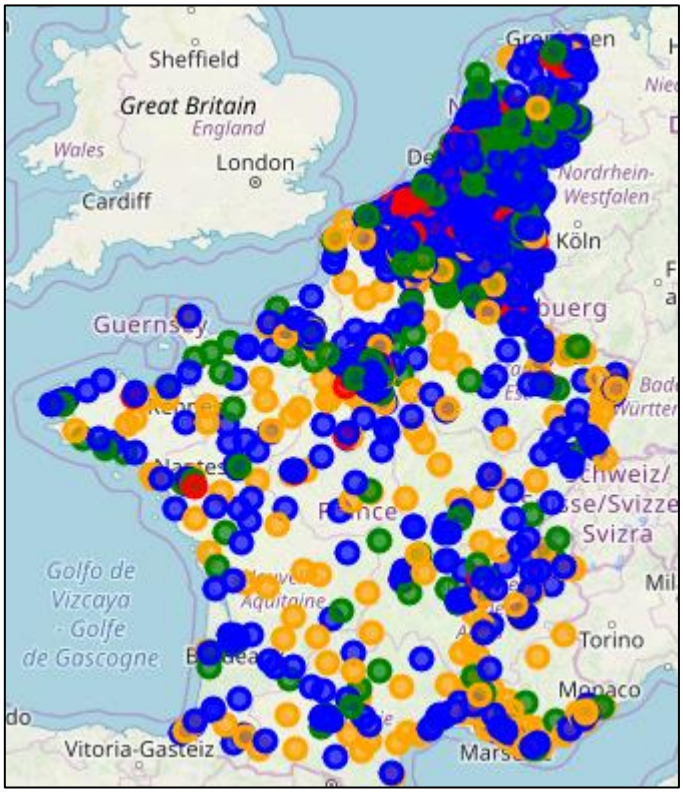


That confirms the observation done in the map, that the orange cluster is by far the most represented in the country, with 85% in Wallonia and Brussels and 90% in Flanders. Looking at it more closely some differences can be spotted like the fact that Flanders seems to be a bit mor homogenous, Brussels seems to have the bigger proportion of the blue cluster and Wallonia the bigger proportion of the red cluster.

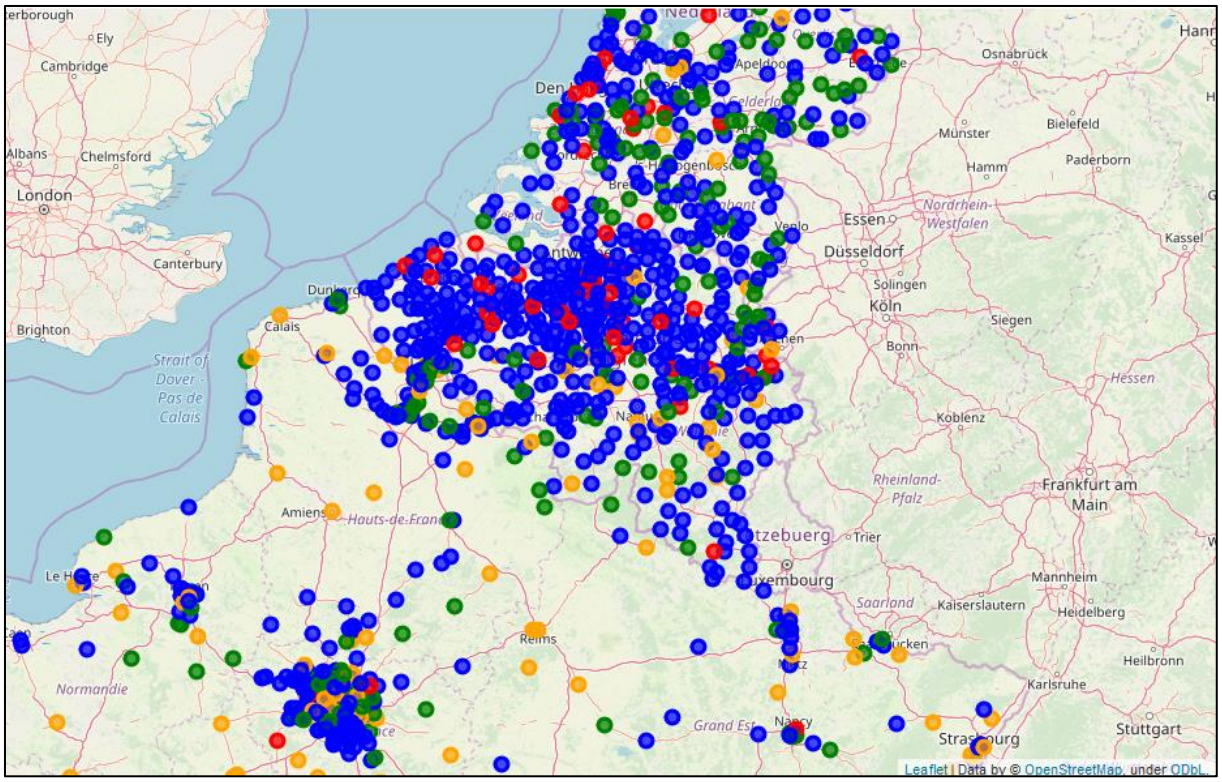
All in all the results does not seems to indicate a big difference between the regions.

Let's try adding France and The Netherlands to the analysis to see what happens.

As for Belgium alone I'll first show the map (first totally zoomed out):

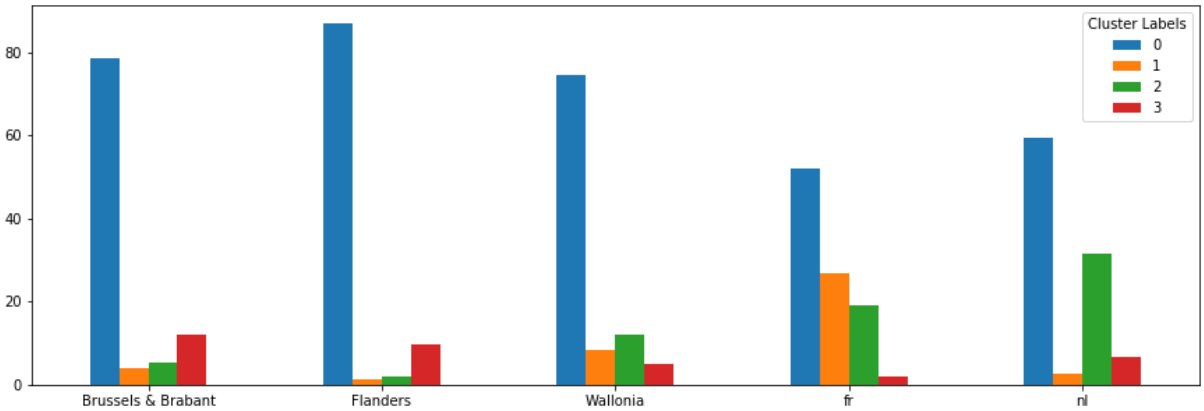


And a bit more zoomed in around Belgium:



As you can notice it is very difficult to make sense of the map here, so let's jump right in the table and bar chart:

NewRegion Cluster Labels	Brussels & Brabant	Flanders	Wallonia	fr	nl
0	78.67	86.92	74.36	51.92	59.28
1	4.00	1.40	8.33	26.92	2.61
2	5.33	1.87	12.18	19.11	31.60
3	12.00	9.81	5.13	2.04	6.51



We see different results with this analysis. First France and The Netherlands and much less homogenous than Belgium's region with only 52% and 59% of their city being part of the bigger cluster.

Their addition also created a new dynamic in Belgium cities that have a drop in their top percentage, especially for Wallonia.

Flanders is once more the one with the more homogenous of all entities and being overall very similar to Brussels.

France and The Netherlands are very similar in term of top percentage, but their second cluster is very different being 27% orange for France and 32% green for the Netherlands.

All in all, Belgium's regions are closer to each other than they are the neighboring countries.

5. Conclusion

The different analysis done in the report show that a lot of cities in Belgium are very similar to one another, and that the difference between different regions are not huge.

They do however exist and can be see both by the percentage of the top cluster and the repartition between the other.

When adding The Netherlands and France, we observed that the difference compared to there countries is much more visible that the one between regions.

There are however still a majority of cities that are similar across the 3 country but the differences between The Netherlands and France show that they are clear differences between countries within the EU.

For initial problem, I believe this bring some data-driven insights to people interested in the topic and can be used as an argument when discussing the issue. It is however only on of the facet of the topic and conclusions should not be made based only on those result.

There are still a lot of information that could be extracted like trying to describe what each of the cluster is, what are the underlying causes that are pushing a city in one direction or the other, a bigger analysis of Europe or even the world to see if new pattern emerge and so on. But those question are outside of the scope of this capstone.

Thanks for reading.