

Social and Cultural Differences between Regions in Belgium

Coursera – IBM Data Science – Capstone project

Table of Contents

1	Introduction	3
1.1	Business understanding	3
1.2	Approach	3
2	Data - Data acquisition	4
2.1	Belgian cities data	4
2.2	Location of the cities	4
2.3	Names of the cities in the three country-languages	5
2.4	City venues data	6
2.5	Geojson file containing the Belgian city borders	8
3	Data - Data preparation and cleansing	9
3.1	The cities dataset	9
3.1.1	Cleansing the cities dataset	9
3.1.2	Dealing with multiple languages	9
3.1.3	Creating the city names to match with the name in the geojson file required by the choropleth maps	10
3.2	The venues dataset	11
3.2.1	Multi-level categories in the venues dataset	11
3.2.2	Preparing the venues dataset for the clustering algorithm¶	12
4	Methodology - Data insight & visualisation	14
4.1	The cities dataset	14
4.2	The venues dataset	17
5	Methodology - Building the analytical model	19
5.1	What machine learning algorithm do we use and why?	19
5.2	Introducing to k-means clustering	19
5.3	Applying k-means clustering	21
5.4	Adding the cluster label to the cities and venues datasets	21
6	Results - Visualizing the results	23
7	Discussion	26
8	Conclusions	27

1 Introduction

1.1 Business understanding

Belgium is a small country in central Europe with approx. 11 million habitants. The capital of Belgium is Brussels, which is also the capital of Europe.

Despite having merely the size of a big city, Belgium is a divided country. There are three communities: a French-speaking community in the South (approx. 4 million), a Dutch-speaking community in the North (approx. 7 million), and a very small German-speaking community.

The fact that there is a political difference is proven with each election. Most people also live with the idea that there are also social and cultural differences, but hard proof is not immediately available.

Many companies – especially the multi-nationals located in Brussels – are working with a mix of employees from both parts of the country. Those companies already found a way to deal with the language difference, but the HR departments of those companies would also want to work with the cultural and differences – if they exist.

The hypothesis that we want to validate:

There are important social and cultural differences between northern (Dutch-speaking) and southern (French-speaking) cities in Belgium.

1.2 Approach

We believe that a natural way to characterize a city - and the people that live in that city - is by the **popularity of its venues**. For example by tallying the amount of parks, bars, restaurants or universities it has relative to all other types of venues, one can get a sense of the cultural and social character of a city.

Therefore, if we could lay our hands on data w.r.t. what the popular venues are in each city, we could use **clustering techniques** to classify cities into categories. These categories can then be visualised on a map to get an idea about the **geographical dispersion** of the categories. If there is a difference between categories mainly appearing in the south and categories mainly appearing in the north, we have proven our hypothesis.

2 Data - Data acquisition

2.1 Belgian cities data










The first dataset that we use is a table that is published on the Dutch Wikipedia that gives an overview of the cities in Belgium with some key metadata.

https://nl.wikipedia.org/wiki/Tabel_van_Belgische_gemeenten

The columns of this table that we are using:

- Name of the city (Gemeente)
- Number of habitants in 2019 (2019 index)
- Acreage of the city (Opp. km²)
- Habitants per km² (inw. per km²)
- Prosperity index (2016 Welv.-index)
- The province to which the city belongs (Provincie of gewest)

Tabel [\[bewerken \]](#)

	Gemeente	1846 Inwoners	1900 Inwoners	1947 Inwoners	2000 Inwoners	2019 Inwoners	1846 index	1900 index	1947 index	2000 index	2019 index	Opp. km ²	Inw. per km ²	2016 Welv.- index	Provincie of gewest
1	Antwerpen	118.682	383.557	526.396	446.525	525.935	100	323	444	376	443	204,51	2.572	88,2	 Antwerpen
2	Gent	128.828	222.895	254.216	224.180	262.219	100	173	197	174	204	156,18	1.679	102,3	 Oost-Vlaanderen
3	Charleroi	54.694	198.837	233.737	200.827	202.267	100	364	427	367	370	102,08	1.982	73,0	 Henegouwen
4	Luik	89.943	206.384	231.502	185.639	197.327	100	229	257	206	219	69,39	2.844	81,4	 Luik
5	Brussel	129.680	218.623	184.838	133.859	181.726	100	169	143	103	140	32,61	5.573	70,0	 Brussel
6	Schaarbeek	6.211	63.508	123.671	105.692	133.309	100	1.023	1.991	1.702	2.146	8,14	16.377	65,4	 Brussel
7	Anderlecht	5.966	47.929	86.412	87.812	119.714	100	803	1.448	1.472	2.007	17,74	6.747	63,8	 Brussel
8	Brugge	60.855	70.277	93.062	116.264	118.325	100	115	153	191	194	138,40	855	111,7	 West-Vlaanderen
9	Namen	42.529	62.427	77.584	105.419	110.779	100	147	182	248	260	175,69	631	99,9	 Namen

This HTML table is parsed using the BeautifulSoup4 Python package, which is aimed at 'screen-scraping' online HTML documents. The next figure illustrates the resulting pandas dataframe.

```
In [11]: df_cities.head()
```

```
Out[11]:
```

	ID	Name	Habitants	Index	Acreage	HabitantsPerSquareKm	ProsperityIndex	Province
0	1	Antwerpen	525.935	443	204,51	2.572	88,2	Antwerpen
1	2	Gent	262.219	204	156,18	1.679	102,3	Oost-Vlaanderen
2	3	Charleroi	202.267	370	102,08	1.982	73,0	Henegouwen
3	4	Luik	197.327	219	69,39	2.844	81,4	Luik
4	5	Brussel	181.726	140	32,61	5.573	70,0	Brussel

Remark: The Name field contains the Dutch name of the city!

2.2 Location of the cities

In order to visualize the cities on a map of Belgium we need their geo-location (latitude-longitude). This location can be retrieved using the geopy package in Python. The resulting pandas dataframe is shown in the following picture.

Social and Cultural Differences between Regions in Belgium

	ID	Name	Habitants	Index	Acreage	HabitantsPerSquareKm	ProsperityIndex	Province	Latitude	Longitude	BoundingBox
0	1	Antwerpen	525.935	443	204,51	2.572	88,2	Antwerpen	51.2211097	4.3997081	[51.1432868, 51.3776412, 4.2175769, 4.4979684]
1	2	Gent	262.219	204	156,18	1.679	102,3	Oost-Vlaanderen	51.0538286	3.7250121	[50.9795422, 51.187946, 3.5797616, 3.849325]
2	3	Charleroi	202.267	370	102,08	1.982	73,0	Henegouwen	50.4120332	4.4436244	[50.3527894, 50.4925149, 4.3474458, 4.5075571]
3	4	Luik	197.327	219	69,39	2.844	81,4	Luik	50.6451381	5.5734203	[50.5610182, 50.6881981, 5.5233883, 5.675257]
4	5	Brussel	181.726	140	32,61	5.573	70,0	Brussel	50.8465573	4.351697	[50.6865573, 51.0065573, 4.191697, 4.511697]

2.3 Names of the cities in the three country-languages

Many cities in Belgium have different names in the different languages. The following file that is published by the federal government contains the city names for the three languages.

<https://economie.fgov.be/sites/default/files/Files/Entreprises/KBO/KBO-codes-identificatie.xls>

	A	B	C	D	E	F
1		NIS GEMEENTECODES / CODES COMMUNES INS				
2	CODE	FRANSE BENAMING	NEDERLANDSE BENAMING	DUITSE BENAMING	BEGINDATUM	EINDDATUM
3	12041		Puurs-Sint-Amands		01.01.2019	31.12.9999
4	44083		Deinze		01.01.2019	31.12.9999
5	44084		Aalter		01.01.2019	31.12.9999
6	44085		Lievegem		01.01.2019	31.12.9999
7	45068		Kruisem		01.01.2019	31.12.9999
8	51067	Enghien	Edingen		01.01.2019	31.12.9999
9	51068	Silly	Opzullik		01.01.2019	31.12.9999
10	51069	Lessines	Lessen		01.01.2019	31.12.9999
11	55085	Seneffe	-	-	01.01.2019	31.12.9999
12	55086	Manage	-	-	01.01.2019	31.12.9999
13	57096	Mouscron	Moeskroen		01.01.2019	31.12.9999
14	57097	Comines-Warneton	Komen-Waasten		01.01.2019	31.12.9999
15	58001	La Louvière	-	-	01.01.2019	31.12.9999
16	58002	Binche	-	-	01.01.2019	31.12.9999
17	58003	Estinnes	-	-	01.01.2019	31.12.9999
18	58004	Morlanwelz	-	-	01.01.2019	31.12.9999
19	72042	-	Oudsbergen	-	01.01.2019	31.12.9999
20	72043		Pelt		01.01.2019	31.12.9999

This file can easily be read and stored in a dataframe using standard pandas features (pd.read_excel). This is the resulting pandas dataframe:

	Frans	Nederlands	Duits	Begin	Einde
CODE					
12041	NaN	Puurs-Sint-Amands	NaN	01.01.2019	31.12.9999
44083	NaN	Deinze	NaN	01.01.2019	31.12.9999
44084	NaN	Aalter	NaN	01.01.2019	31.12.9999
44085	NaN	Lievegem	NaN	01.01.2019	31.12.9999
45068	NaN	Kruisem	NaN	01.01.2019	31.12.9999
51067	Enghien	Edingen	NaN	01.01.2019	31.12.9999
51068	Silly	Opzullik	NaN	01.01.2019	31.12.9999
51069	Lessines	Lessen	NaN	01.01.2019	31.12.9999
55085	Seneffe	-	-	01.01.2019	31.12.9999
55086	Manage	-	-	01.01.2019	31.12.9999

2.4 City venues data

A good source of information for venues all over the world is Foursquare. Foursquare offers a Places API to gain real-time access to Foursquare's global database of rich venue data and user content. One can easily find popular venues in a city or a location by using the explore function in the Place API.

Remark: You have to register to get a client_id and secret to be able to access that API. By using the free subscription you can make a limited number of API calls per day.

GET <https://api.foursquare.com/v2/venues/explore>

You can simply pass the following URL parameters

client_id and client_secret	These are your credentials that you receive when you subscribe to the API and that you need to pass with each API call.
v	The version of the API that you want to access. e.g. '20180605'
near	The location that you want the venues for. e.g. 'Lokeren, Belgium'
radius	Radius to search within, in meters. If radius is not specified, a suggested radius will be used based on the density of venues in the area. e.g. '2000' (2 km's) Based on some experimentation, we calculated the radius from the city acreage using the following formula: $\text{radius} = \text{acreage} / 70 \times 1000$
limit	Number of results to return, up to 100.

and what you get back as a result is a list of venues in a JSON format.

The following extract of a result message gives you an idea about the venue data this is returned.

```
{'reasons': {'count': 0,  
  'items': [{'summary': 'This spot is popular',
```

Social and Cultural Differences between Regions in Belgium

```
{
  'type': 'general',
  'reasonName': 'globalInteractionReason'}],
'venue': {'id': '4c5093d9bd099521bed1525e',
'name': 'De Donkere Wolk',
'location': {'address': 'Torenstraat 14',
'lat': 51.105140425790516,
'lng': 3.991317566213707,
'labeledLatLngs': [{'label': 'display',
'lat': 51.105140425790516,
'lng': 3.991317566213707}]},
'postalCode': '9160',
'cc': 'BE',
'city': 'Lokeren',
'state': 'Oost-Vlaanderen',
'country': 'België',
'formattedAddress': ['Torenstraat 14', '9160 Lokeren', 'België']],
'categories': [{'id': '4bf58dd8d48988d116941735',
'name': 'Bar',
'pluralName': 'Bars',
'shortName': 'Bar',
'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/nightlife/pub_',
'suffix': '.png'},
'primary': True}],
'photos': {'count': 0, 'groups': []},
'referralId': 'e-0-4c5093d9bd099521bed1525e-0'}
```

This JSON file can easily be parsed and stored in Python dictionary data structure. The data from that dictionary can then be stored in a pandas dataframe.

We can do that for each city in the dataframe that holds the complete list of all Belgian cities (see section 2.1). The resulting dataframe looks like:

```
In [31]: be_venues.head(10)
```

```
Out[31]:
```

	City	Venue	Venue Latitude	Venue Longitude	Category Class	Venue Category
0	Antwerpen	Moochie Frozen Yoghurt	51.220036	4.402850	https://ss3.4sqi.net/img/categories_v2/food/fr...	Frozen Yogurt Shop
1	Antwerpen	Dogma Cocktails	51.221146	4.402854	https://ss3.4sqi.net/img/categories_v2/nightli...	Cocktail Bar
2	Antwerpen	Absinthbar	51.219912	4.400709	https://ss3.4sqi.net/img/categories_v2/nightli...	Cocktail Bar
3	Antwerpen	Pitten en Bonen	51.217657	4.402712	https://ss3.4sqi.net/img/categories_v2/food/ju...	Juice Bar
4	Antwerpen	Kartini Indonesisch Restaurant	51.219270	4.400557	https://ss3.4sqi.net/img/categories_v2/food/in...	Indonesian Restaurant
5	Antwerpen	Hunkemöller	51.218611	4.405531	https://ss3.4sqi.net/img/categories_v2/shops/a...	Lingerie Store
6	Antwerpen	Brasserie Appelmans	51.219879	4.400717	https://ss3.4sqi.net/img/categories_v2/nightli...	Cocktail Bar
7	Antwerpen	Quetzal	51.220625	4.402132	https://ss3.4sqi.net/img/categories_v2/food/co...	Coffee Shop
8	Antwerpen	Maison Tartine	51.221703	4.404996	https://ss3.4sqi.net/img/categories_v2/food/deli...	Sandwich Place
9	Antwerpen	Bia Mara	51.220894	4.400189	https://ss3.4sqi.net/img/categories_v2/food/fi...	Fish & Chips Shop

2.5 Geojson file containing the Belgian city borders

In order to visualize our results in a Folium [choropleth](#), we need a geo-json file containing the location data for the city borders. We found one in GitHub. It's not a very recent one, but it will do the trick for this project.

<https://github.com/Datafable/rolling-blackout-belgium/blob/master/data/geospatial/municipalities-belgium.geojson>

Remark: For cities that are officially bilingual a somewhat special format is used in this file:

- dutchname#frenchname when the primary language is Dutch - e.g. Brussel#Bruxelles
- frenchname#dutchname when the primary language is French - e.g. Enghien#Edingen

We need to have that special name in our input dataset, so that the city can be mapped on the location in the geojson file.

3 Data - Data preparation and cleansing

3.1 The cities dataset

3.1.1 Cleansing the cities dataset

In the cities dataset that was scraped from the Dutch Wikipedia, the numbers are formatted “Belgian style” (e.g. commas as decimal separator, periods as thousands group separator).

For the **prosperity index** and the **acreage** data fields, commas were replaced by periods for the decimal separator.

For the **habitants** and **habitants per km²** data fields, the thousands group separator periods were removed.

The fields mentioned above were still of type ‘string’ (object) in the original dataframe. Those fields were translated to numeric values.

```
ID                int64
Name              object
Habitants         int64
Index            float64
Acreage           float64
HabitantsPerSquareKm  int64
ProsperityIndex   float64
Province          object
Latitude          float64
Longitude         float64
BoundingBox       object
dtype: object
```

The resulting dataframe looks like:

```
Out[11]:
```

	ID	Name	Habitants	Index	Acreage	HabitantsPerSquareKm	ProsperityIndex	Province	Latitude	Longitude	BoundingBox
0	1	Antwerpen	525935	443.0	204.51	2572	88.2	Antwerpen	51.221110	4.399708	['51.1432868', '51.3776412', '4.2175769', '4.4...
1	2	Gent	262219	204.0	156.18	1679	102.3	Oost-Vlaanderen	51.053829	3.725012	['50.9795422', '51.187946', '3.5797616', '3.84...
2	3	Charleroi	202267	370.0	102.08	1982	73.0	Henegouwen	50.412033	4.443624	['50.3527894', '50.4925149', '4.3474458', '4.5...
3	4	Luik	197327	219.0	69.39	2844	81.4	Luik	50.645138	5.573420	['50.5610182', '50.6881981', '5.5233883', '5.6...
4	5	Brussel	181726	140.0	32.61	5573	70.0	Brussel	50.846557	4.351697	['50.6865573', '51.0065573', '4.191697', '4.51...

3.1.2 Dealing with multiple languages

As mentioned before, many cities in Belgium have different names in the different languages. First of all we made sure that we have a proper dataset of all the cities containing all the different languages.

The multi-language cities dataset contains a full **historic view** on the city names. The rows have a start and an end date. We only want the most recent version of the names for each city, so **outdated rows were removed** from the dataset.

Then, a **new Name column** was created to merge the multi-language dataset with the cities dataset mentioned in the previous section. The cities in that dataset use the Dutch name if one exists and

Social and Cultural Differences between Regions in Belgium

the French name if one doesn't. This column is then used to match the cities with the rows in the cities dataset.

When merged, all rows in the cities dataframe – the dataset with the key characteristics of the cities – could be matched.

```
Cities in original cities dataframe 581
Cities in original multi-language dataframe 636
Cities in merged dataframe 581
```

This is an extract of the resulting dataframe after merging the cities and the multi-language datasets:

Out[31]:

	ID	Name	Habitants	Acreage	Frans	Nederlands	Duits
0	1	Antwerpen	525935	204.51	Anvers	Antwerpen	Antwerpen
1	2	Gent	262219	156.18	Gand	Gent	Gent
2	3	Charleroi	202267	102.08	Charleroi	-	-
3	4	Luik	197327	69.39	Liège	Luik	Lüttich
4	5	Brussel	181726	32.61	Bruxelles	Brussel	-
5	6	Schaerbeek	133309	8.14	Schaerbeek	Schaerbeek	-
6	7	Anderlecht	119714	17.74	Anderlecht	Anderlecht	-
7	8	Brugge	118325	138.40	Bruges	Brugge	Brügge
8	9	Namen	110779	175.69	Namur	Namen	-
9	10	Leuven	101624	56.63	Louvain	Leuven	Löwen

Remark: the view above does not contain all columns and rows of the new dataframe!

3.1.3 Creating the city names to match with the name in the geojson file required by the choropleth maps

As explained in the previous chapter, the geojson file uses a special format for the city names. For most of the cities this name can be derived from the dataset specified above using the following rule:

“If the city is in a French-speaking province use French name, else use Dutch name.”

For the cities in Brussels district, which has a special status, the geojson name consists of 2 parts separated by a #. This is also the case for some other cities that have a Dutch and a French name.

There were some other inconsistencies that were solved manually (e.g. cities that recently changed names after merging with another city).

The following extract of the cities dataset shows the result (geojson_name in the last column):

Social and Cultural Differences between Regions in Belgium

Out[33]:

SquareKm	ProsperityIndex	Province	Latitude	Longitude	BoundingBox	Frans	Nederlands	Duits	Begin	Einde	geojson_name
2572	88.2	Antwerpen	51.221110	4.399708	['51.1432868', '51.3776412', '4.2175765', '4.4...	Anvers	Antwerpen	Antwerpen	01.01.0001	31.12.9999	Antwerpen
1679	102.3	Oost-Vlaanderen	51.053829	3.725012	['50.9795422', '51.187946', '3.5797616', '3.84...	Gand	Gent	Gent	01.01.0001	31.12.9999	Gent
1982	73.0	Henegouwen	50.412033	4.443624	['50.3527894', '50.4925149', '4.3474458', '4.5...	Charleroi	-	-	01.01.0001	31.12.9999	Charleroi
2844	81.4	Luik	50.645138	5.573420	['50.5610182', '50.6881981', '5.5233883', '5.6...	Liège	Luik	Lüttich	01.01.0001	31.12.9999	Liège
5573	70.0	Brussel	50.846557	4.351697	['50.6865573', '51.0065573', '4.191697', '4.51...	Bruxelles	Brussel	-	01.01.0001	31.12.9999	Brussel#Bruxelles
16377	65.4	Brussel	50.867604	4.373712	['50.8434069', '50.8811977', '4.3571322', '4.4...	Schaerbeek	Schaerbeek	-	01.01.0001	31.12.9999	Schaerbeek#Schaerbeek
6747	63.8	Brussel	50.839098	4.329653	['50.8070598', '50.850549', '4.2437658', '4.34...	Anderlecht	Anderlecht	-	01.01.0001	31.12.9999	Anderlecht
855	111.7	West-Vlaanderen	51.208553	3.226772	['51.1581918', '51.363347', '3.1341802', '3.30...	Bruges	Brugge	Brügge	01.01.0001	31.12.9999	Brugge
631	99.9	Namen	50.466528	4.866189	['50.3872825', '50.5313007', '4.723053', '4.98...	Namur	Namen	-	01.01.0001	31.12.9999	Namur
1794	113.6	Vlaams-Brabant	50.879202	4.701168	['50.8242096', '50.9440707', '4.640295', '4.77...	Louvain	Leuven	Löwen	01.01.0001	31.12.9999	Leuven

3.2 The venues dataset

3.2.1 Multi-level categories in the venues dataset

Foursquare uses a max 3-level category classification tree to assign categories to venues. This is only visible in the icon prefix URL (see below).

```
'categories': [{ 'id': '4bf58dd8d48988d116941735',  
  'name': 'Bar',  
  'pluralName': 'Bars',  
  'shortName': 'Bar',  
  'icon': { 'prefix': 'https://ss3.4sqi.net/img/categories_v2/nightlife/pub',  
    'suffix': '.png'},  
  'primary': True}],
```

For the example above the 3 levels are:

- Level 1: nightlife
- Level 2: pub
- Level 3: Bar

We have parsed the URL and added two extra columns to the original dataset:

- **class1** for the level 1 category
- **class2** for the level 2 category
- (The Category Class column already contained the level 3 category)

Social and Cultural Differences between Regions in Belgium

Out[20]:

	City	Venue	Venue Latitude	Venue Longitude	Category Class	Venue Category	class1	class2
0	Antwerpen	Moochie Frozen Yoghurt	51.220036	4.402850	https://ss3.4sqi.net/img/categories_v2/food/fr...	Frozen Yogurt Shop	food	food frozenyogurt_
1	Antwerpen	Dogma Cocktails	51.221146	4.402854	https://ss3.4sqi.net/img/categories_v2/nightli...	Cocktail Bar	nightlife	nightlife cocktails_
2	Antwerpen	Absinthbar	51.219912	4.400709	https://ss3.4sqi.net/img/categories_v2/nightli...	Cocktail Bar	nightlife	nightlife cocktails_
3	Antwerpen	Pitten en Bonen	51.217657	4.402712	https://ss3.4sqi.net/img/categories_v2/food/ju...	Juice Bar	food	food juicebar_
4	Antwerpen	Kartini Indonesisch Restaurant	51.219270	4.400557	https://ss3.4sqi.net/img/categories_v2/food/in...	Indonesian Restaurant	food	food indonesian_
5	Antwerpen	Hunkemöller	51.218611	4.405531	https://ss3.4sqi.net/img/categories_v2/shops/a...	Lingerie Store	shops	shops apparel_lingerie_
6	Antwerpen	Brasserie Appelmans	51.219879	4.400717	https://ss3.4sqi.net/img/categories_v2/nightli...	Cocktail Bar	nightlife	nightlife cocktails_
7	Antwerpen	Quetzal	51.220625	4.402132	https://ss3.4sqi.net/img/categories_v2/food/co...	Coffee Shop	food	food coffeeshop_
8	Antwerpen	Maison Tartine	51.221703	4.404996	https://ss3.4sqi.net/img/categories_v2/food/deli...	Sandwich Place	food	food deli_
9	Antwerpen	Bia Mara	51.220894	4.400189	https://ss3.4sqi.net/img/categories_v2/food/fl...	Fish & Chips Shop	food	food fishandchips_

3.2.2 Preparing the venues dataset for the clustering algorithm

The goal of the clustering technique that we will describe in the next chapter is to group cities with similar characteristics w.r.t. popular venues.

What we will describe in this section is how we can create a dataset with the 5 most popular venues for each city, based on the venues dataset that we used so far.

The first step is to use the onehot encoding technique. One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. The categorical variable in our case is Venue Category column in the venues dataset shown above. The result is a dataset that looks like the following one:

Out[5]:

	City	ATM	Accessories Store	African Restaurant	Airport	Airport Service	American Restaurant	Amphitheater	Antique Shop	Apres Ski Bar	...	Weight Loss Center	Whisky Bar	Windmill	Wine Bar	Wine Shop	Winery
0	Antwerpen	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
1	Antwerpen	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
2	Antwerpen	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
3	Antwerpen	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
4	Antwerpen	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0

5 rows x 444 columns

We then use this dataset to calculate the mean of the occurrences for each category for each city. The result looks like:

Out[6]:

	City	ATM	Accessories Store	African Restaurant	Airport	Airport Service	American Restaurant	Amphitheater	Antique Shop	Apres Ski Bar	...	Weight Loss Center	Whisky Bar	Windmill	Wine Bar	Wine Shop	W
0	's-Gravenbrakel	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.000000	0.0	
1	Aalst	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.020000	0.0	
2	Aalter*	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.011905	0.0	
3	Aarlen	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.000000	0.0	
4	Aarschot	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.000000	0.0	

5 rows x 444 columns

This dataset is then translated into the following one, where we have the 5 most popular venue categories for each city.

Social and Cultural Differences between Regions in Belgium

Out[9]:

	City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	's-Gravenbrakel	Platform	Friterie	Chinese Restaurant	Stadium	Asian Restaurant
1	Aalst	Bar	Coffee Shop	Clothing Store	Belgian Restaurant	Bistro
2	Aalter*	Bar	Bakery	Supermarket	Belgian Restaurant	Friterie
3	Aarlen	Bar	Italian Restaurant	Supermarket	Burger Joint	Pizza Place
4	Aarschot	Bar	Restaurant	Italian Restaurant	Friterie	Pub
5	Aartselaar	Bar	Asian Restaurant	Weight Loss Center	Lingerie Store	Athletics & Sports
6	Aat	Supermarket	Italian Restaurant	Bar	Clothing Store	Electronics Store
7	Affligem	Rental Car Location	Fishing Spot	Sake Bar	Cocktail Bar	Zoo
8	Aiseau-Presles	Supermarket	Italian Restaurant	Restaurant	Fast Food Restaurant	Clothing Store
9	Alken	Bar	Restaurant	Brasserie	Playground	Brewery
10	Alveringem	Bar	Pub	Restaurant	Pharmacy	Cafeteria
11	Amay	Supermarket	Chinese Restaurant	Restaurant	Basketball Stadium	Basketball Court
12	Amel	Friterie	Supermarket	Restaurant	Soccer Field	Zoo
13	Andenne	Supermarket	Bar	Bakery	Athletics & Sports	Friterie
14	Anderlecht	Bar	Supermarket	Sandwich Place	Italian Restaurant	Sports Bar
15	Anderlues	Bar	Italian Restaurant	Supermarket	Pharmacy	Food
16	Anhée	Supermarket	Stadium	Bakery	Gastropub	Gym / Fitness Center
17	Ans	Chinese Restaurant	Thai Restaurant	Basketball Court	Italian Restaurant	Pizza Place
18	Anthisnes	Museum	Recreation Center	Bistro	Food Court	Fast Food Restaurant
19	Antoing	Chinese Restaurant	Supermarket	Bakery	Historic Site	Sandwich Place

This dataset will be used by the clustering algorithm to find groups of similar cities – cities that have similar most popular venues.

4 Methodology - Data insight & visualisation

4.1 The cities dataset

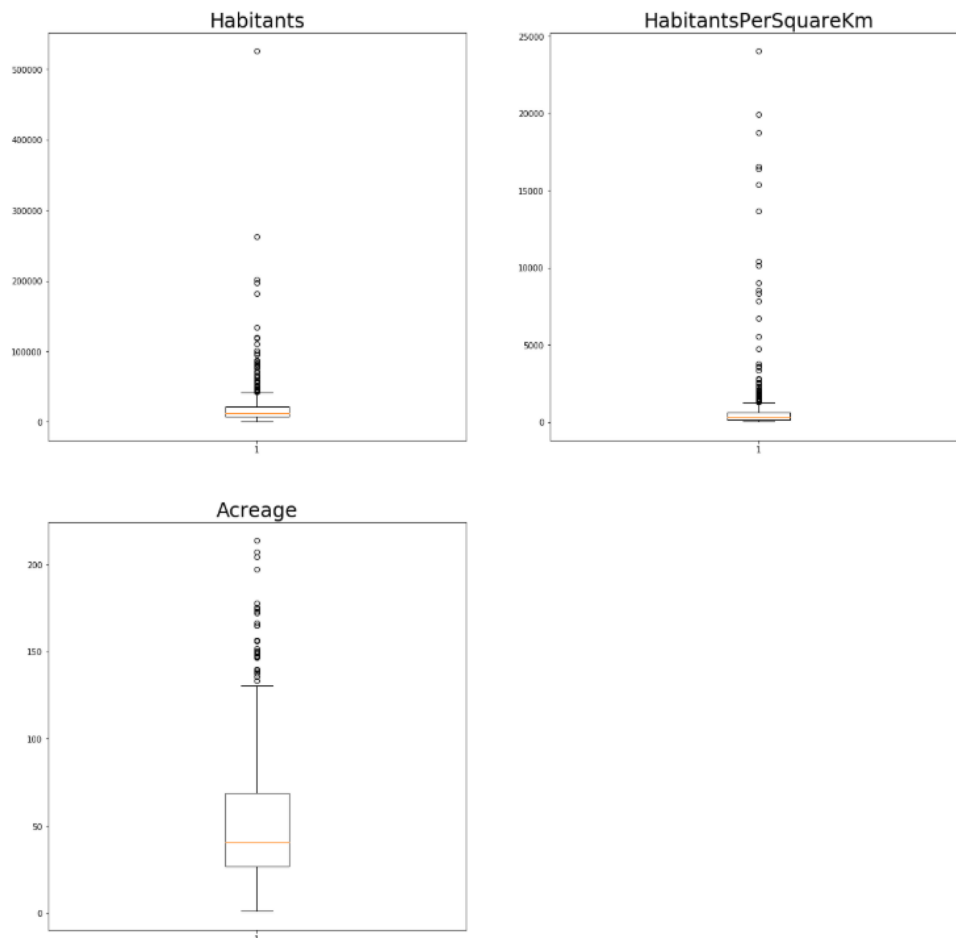
The cities dataset mainly contains numeric values. Some key characteristics of the dataset are summarized in the following table.

```
In [23]: df_cities_and_lang.describe()
```

```
Out[23]:
```

	ID	Habitants	Index	Acreage	HabitantsPerSquareKm	ProsperityIndex	Latitude	Longitude
count	581.000000	581.000000	581.000000	581.000000	581.000000	575.000000	580.000000	580.000000
mean	291.000000	19675.977625	258.604811	52.543769	794.036145	103.553565	50.739602	4.561950
std	167.864529	31727.851917	196.508641	38.250614	2197.952030	14.289263	0.378283	0.833358
min	1.000000	83.000000	1.000000	1.140000	25.000000	49.600000	49.539075	2.580670
25%	146.000000	7357.000000	125.000000	26.920000	165.000000	94.100000	50.521429	4.009953
50%	291.000000	12436.000000	200.000000	40.380000	313.000000	103.500000	50.797974	4.545063
75%	436.000000	21227.000000	340.000000	68.900000	618.000000	112.350000	51.012278	5.243842
max	581.000000	525935.000000	967.000000	213.750000	24037.000000	159.000000	51.467796	6.257827

The following boxplots give us an idea about the distribution of the values for the respective data fields (habitants, habitants per km², acreage):



One can see that the datafields habitants and habitants per km² have a lot of outliers.

Social and Cultural Differences between Regions in Belgium

For the habitants:

```
Number of outliers: 145
Minimum number of habitants: 21227.0
Maximum number of habitants: 525935
```

Out[38]:

	ID	Name	Habitants	HabitantsPerSquareKm	Acreage
0	1	Antwerpen	525935	2572	204.51
1	2	Gent	262219	1679	156.18
2	3	Charleroi	202267	1982	102.08
3	4	Luik	197327	2844	69.39
4	5	Brussel	181726	5573	32.61
5	6	Schaarbeek	133309	16377	8.14
6	7	Anderlecht	119714	6747	17.74
7	8	Brugge	118325	855	138.40
8	9	Namen	110779	631	175.69
9	10	Leuven	101624	1794	56.63
10	11	Sint-Jans-Molenbeek	97462	16542	5.89
11	12	Bergen	95613	653	146.53
12	13	Elsene	86876	13693	6.34
13	14	Mechelen	86616	1329	65.19
14	15	Aalst	86445	1107	78.12
15	16	Ukkel	83024	3624	22.91
16	17	La Louvière	80757	1257	64.24

For the habitants per km²:

```
Number of outliers: 145
Minimum number of habitants per square km: 618.0
Maximum number of habitants per square km: 24037
```

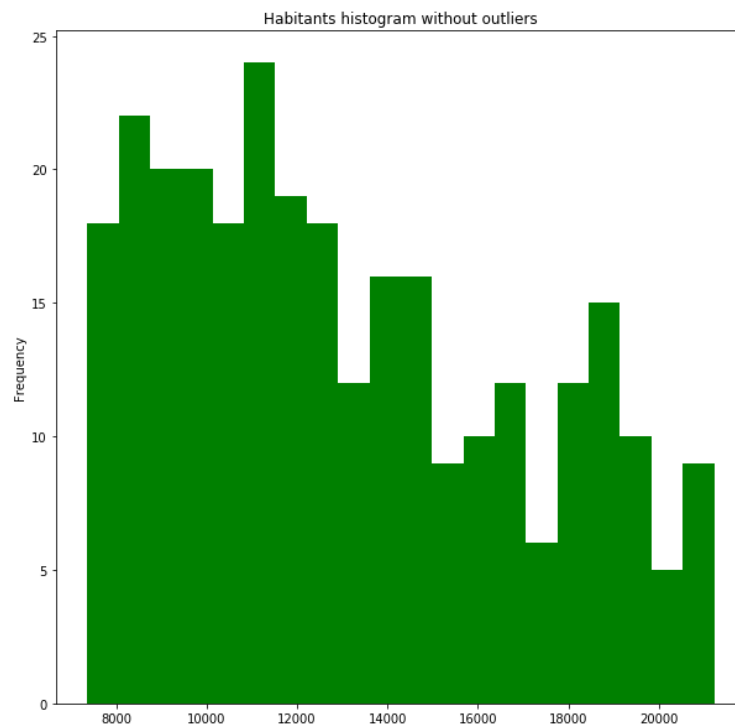
Out[39]:

	ID	Name	Habitants	HabitantsPerSquareKm	Acreage
93	94	Sint-Joost-ten-Node	27457	24037	1.14
30	31	Sint-Gillis	50267	19910	2.52
136	137	Koekelberg	21990	18755	1.17
10	11	Sint-Jans-Molenbeek	97462	16542	5.89
5	6	Schaarbeek	133309	16377	8.14
32	33	Etterbeek	48367	15358	3.15
12	13	Elsene	86876	13693	6.34
29	30	Jette	52536	10417	5.04
114	115	Ganshoren	24902	10142	2.46
27	28	Vorst	56289	9009	6.25
112	113	Sint-Agatha-Berchem	25179	8537	2.95
41	42	Evere	41763	8322	5.02
26	27	Sint-Lambrechts-Woluve	56660	7842	7.22
6	7	Anderlecht	119714	6747	17.74
4	5	Brussel	181726	5573	32.61
40	41	Sint-Pieters-Woluve	41824	4725	8.85
64	65	Oudergem	34013	3765	9.03
15	16	Ukkel	83024	3624	22.91

Social and Cultural Differences between Regions in Belgium

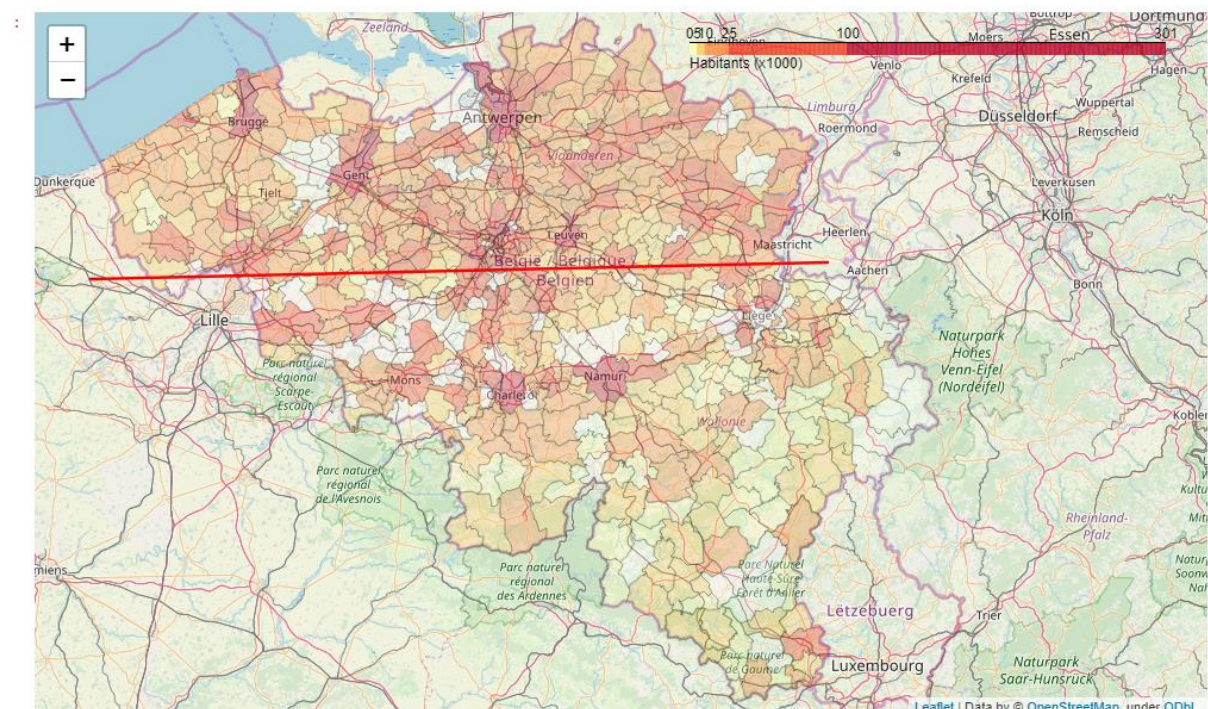
Remark: it is not a surprise that these are all cities in the Brussels district.

The following histogram shows the distribution of habitants per city when you exclude the outliers. Most of cities in Belgium have between 8000 and 20000 habitants.



The following choropleth shows the habitants per city on a map of Belgium. The line on the map represents the language border between the north and the south (in reality this is not a straight line, but instead follows the city borders!).

Remark: the cities filled with a white colour on the map are cities for which data is missing in the cities dataset.



This map shows that in general the cities in the north have higher populations than the cities in the south.

4.2 The venues dataset

There are 17553 venues in this dataset.

There are 9 unique main (level 1) categories.

```
['food' 'nightlife' 'shops' 'parks_outdoors' 'building'
 'arts_entertainment' 'travel' 'event' 'education']
```

There are 261 unique level 2 categories.

```
['food frozenyogurt_' 'nightlife cocktails_' 'food juicebar_'
 'food indonesian_' 'shops apparel lingerie_' 'food coffeeshop_'
 'food deli_' 'food fishandchips_' 'nightlife pub_' 'food sushi_'
 'parks outdoors plaza_' 'building religious church_' 'shops apparel_'
 'food cupcakes_' 'shops food_butcher_' 'shops apparel_women_'
 'food falafel_' 'shops apparel_boutique_' 'food italian_' 'food default_'
 'arts_entertainment musicvenue_jazzclub_' ...]
```

There are 443 unique categories.

```
['Frozen Yogurt Shop' 'Cocktail Bar' 'Juice Bar' 'Indonesian Restaurant'
 'Lingerie Store' 'Coffee Shop' 'Sandwich Place' 'Fish & Chips Shop'
 'Beer Bar' 'Sushi Restaurant' 'Plaza' 'Church' 'Clothing Store' 'Bar'
 'Cupcake Shop' 'Kitchen Supply Store' 'Women's Store'
 'Falafel Restaurant' 'Boutique' 'Italian Restaurant' 'Pub' 'Restaurant'
 'Jazz Club' 'Asian Restaurant' 'Soup Place' 'Deli / Bodega'
 'Chocolate Shop' 'Shoe Store' 'Belgian Restaurant' 'Bookstore'
 'Breakfast Spot' 'Spanish Restaurant' 'Donut Shop' 'Road'
 'French Restaurant' 'Tapas Restaurant' 'Optical Shop' ... ]
```

The following extract gives us an idea of the most popular venues in each city:

```
----'s-Gravenbrakel----
      venue  freq
0      Platform 0.15
1      Friterie 0.12
2  Chinese Restaurant 0.08
3          Pool 0.04
4  Asian Restaurant 0.04

----Aalst----
      venue  freq
0          Bar 0.11
1  Coffee Shop 0.07
2  Clothing Store 0.05
3  Belgian Restaurant 0.04
4          Bistro 0.04

----Aalter*----
      venue  freq
0          Bar 0.11
1      Bakery 0.07
2  Supermarket 0.06
```

Social and Cultural Differences between Regions in Belgium

```
3 Belgian Restaurant 0.05
4 Friterie 0.04
```

----Aarlen----

```
venue freq
0 Bar 0.13
1 Italian Restaurant 0.11
2 Supermarket 0.09
3 French Restaurant 0.04
4 Burger Joint 0.04
```

----Aarschot----

```
venue freq
0 Bar 0.08
1 Restaurant 0.06
2 Friterie 0.04
3 Pub 0.04
4 Italian Restaurant 0.04
```

----Aartselaar----

```
venue freq
0 Bar 0.20
1 Bookstore 0.07
2 Lingerie Store 0.07
3 Convenience Store 0.07
4 Plaza 0.07
```

----Aat----

```
venue freq
0 Supermarket 0.18
1 Clothing Store 0.05
2 Electronics Store 0.05
3 Italian Restaurant 0.05
4 Bar 0.05
```

----Affligem----

```
venue freq
0 Rental Car Location 0.25
1 Fishing Spot 0.25
2 Sake Bar 0.25
3 Cocktail Bar 0.25
4 Paintball Field 0.00
```

5 Methodology - Building the analytical model

5.1 What machine learning algorithm do we use and why?

Remember the hypothesis that we want to validate:

There are important social and cultural differences between northern (Dutch-speaking) and southern (French-speaking) cities in Belgium.

So, what we actually want to prove is that there are important dissimilarities between cities in the north and in the south. One popular method to find similarities between entities – for which we have data – are clustering algorithms. A clustering algorithm classifies entities – the cities in our case – based on the data available for the cities.

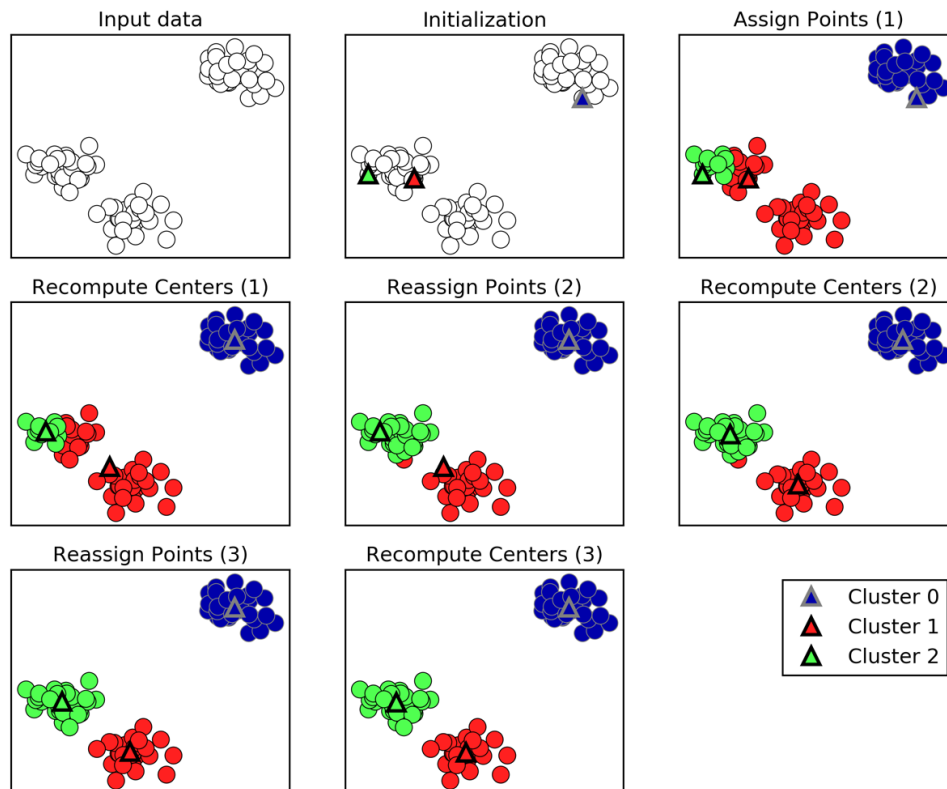
The data that we want to use is the most popular venues for each city (a dataset that we prepared in the previous section). If the results of the algorithm shows that for one cluster most of the cities lie in the north, while for another cluster most of the cities lie in the south we have validated our hypothesis.

We choose to use a simple clustering algorithm, called k-means clustering.

5.2 Introducing to k-means clustering

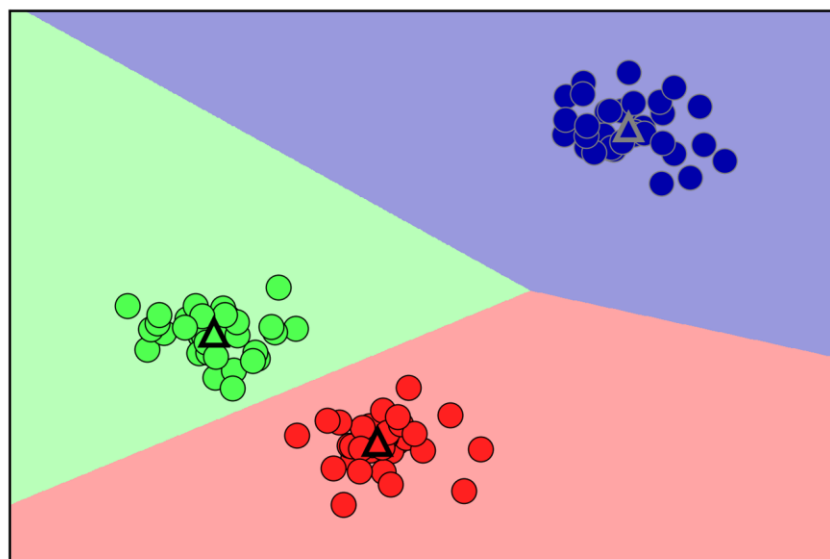
The analytical technique that we selected to classify cities based on the most popular venues is a clustering model called k-means clustering.

k-means clustering is one of the simplest and most commonly used clustering algorithms. It tries to find cluster centers that are representative of certain regions of the data. The algorithm alternates between two steps: assigning each data point to the closest cluster center, and then setting each cluster center as the mean of the data points that are assigned to it. The algorithm is finished when the assignment of instances to clusters no longer changes. The following example illustrates the algorithm on a synthetic dataset:



Cluster centers are shown as triangles, while data points are shown as circles. Colors indicate cluster membership. We specified that we are looking for three clusters, so the algorithm was initialized by declaring three data points randomly as cluster centers. Then the iterative algorithm starts. First, each data point is assigned to the cluster center it is closest to (see “Assign Points (1)”). Next, the cluster centers are updated to be the mean of the assigned points (see “Recompute Centers (1)”). Then the process is repeated two more times. After the third iteration, the assignment of points to cluster centers remained unchanged, so the algorithm stops.

Given new data points, *k*-means will assign each to the closest cluster center. The next example shows the boundaries of the cluster centers that were learned in the previous figure.



5.3 Applying k-means clustering

We used the k-means clustering algorithm implementation available in the sklearn Python package.

Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

After experimentation we saw that using 5 as the number of clusters. The algorithm yields the following results:

```
Number of cat 0 cities: 180
Number of cat 1 cities: 27
Number of cat 2 cities: 17
Number of cat 3 cities: 247
Number of cat 4 cities: 99
```

5.4 Adding the cluster label to the cities and venues datasets

In order to be able to visualize our results and to be able to draw conclusions we added the cluster label for the cities as an extra column in the cities dataset and the venues dataset.

Cities dataset:

	Name	Habitants	Acreage	HabitantsPerSquareKm	Cluster Labels
0	Antwerpen	525935	204.51	2572	3
1	Gent	262219	156.18	1679	3
2	Charleroi	202267	102.08	1982	3
3	Luik	197327	69.39	2844	3
4	Brussel	181726	32.61	5573	3
5	Schaarbeek	133309	8.14	16377	3
6	Anderlecht	119714	17.74	6747	0
7	Brugge	118325	138.40	855	3
8	Namen	110779	175.69	631	3
9	Leuven	101624	56.63	1794	0
10	Sint-Jans-Molenbeek	97462	5.89	16542	4
11	Bergen	95613	146.53	653	3
12	Elsene	86876	6.34	13693	3
13	Mechelen	86616	65.19	1329	3
14	Aalst	86445	78.12	1107	0
15	Ukkel	83024	22.91	3624	3
16	La Louvière	80757	64.24	1257	3
17	Hasselt	78296	102.24	766	3
18	Sint-Niklaas	77769	83.80	927	3
19	Kortrijk	76735	80.02	959	3

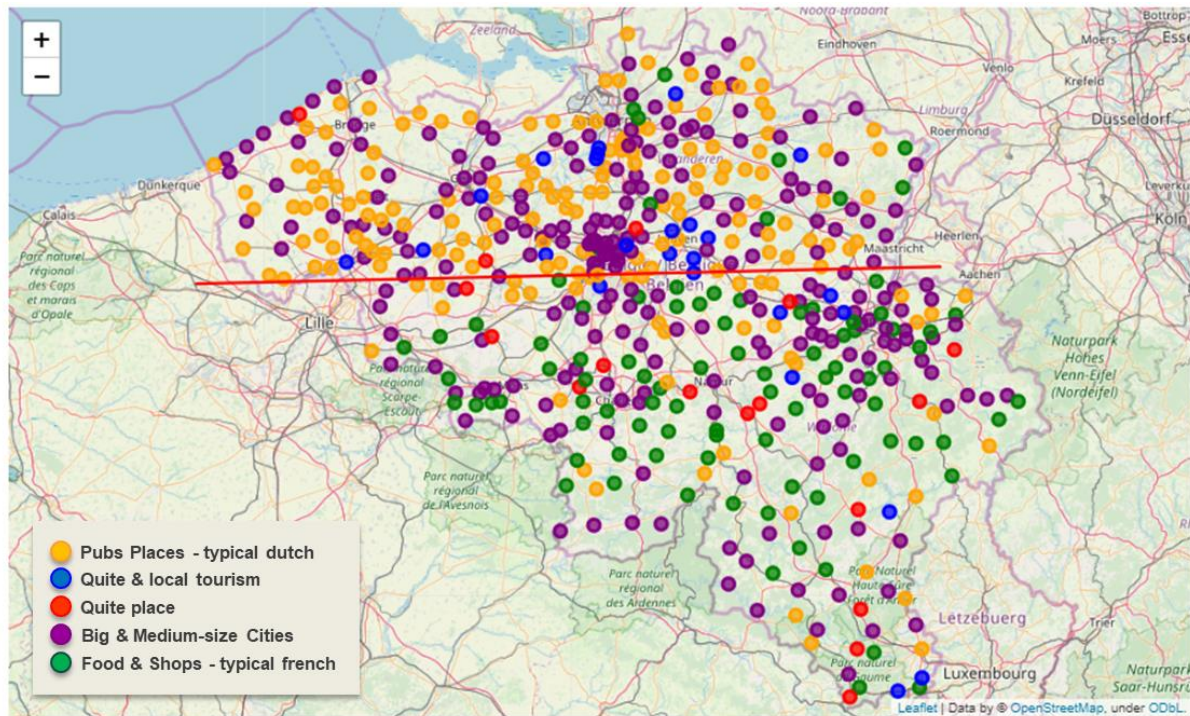
Venues dataset:

Social and Cultural Differences between Regions in Belgium

	City	Venue	Venue Latitude	Venue Longitude	Category Class	Venue Category	class1	class2	Cluster Labels
0	Antwerpen	Moochie Frozen Yoghurt	51.220036	4.402850	https://ss3.4sqi.net/img/categories_v2/food/fr...	Frozen Yogurt Shop	food	food frozenyogurt_	3
1	Antwerpen	Dogma Cocktails	51.221146	4.402854	https://ss3.4sqi.net/img/categories_v2/nightli...	Cocktail Bar	nightlife	nightlife cocktails_	3
2	Antwerpen	Absinthbar	51.219912	4.400709	https://ss3.4sqi.net/img/categories_v2/nightli...	Cocktail Bar	nightlife	nightlife cocktails_	3
3	Antwerpen	Pitten en Bonen	51.217657	4.402712	https://ss3.4sqi.net/img/categories_v2/food/ju...	Juice Bar	food	food juicebar_	3
4	Antwerpen	Kartini Indonesisch Restaurant	51.219270	4.400557	https://ss3.4sqi.net/img/categories_v2/food/in...	Indonesian Restaurant	food	food indonesian_	3

6 Results - Visualizing the results

The following Folium map shows the cities with their associated cluster types on the map of Belgium. We have assigned a meaningful label to each cluster that aims at giving a general description of the types of cities in that cluster. Those labels and associated colors can be found in the map's legend.



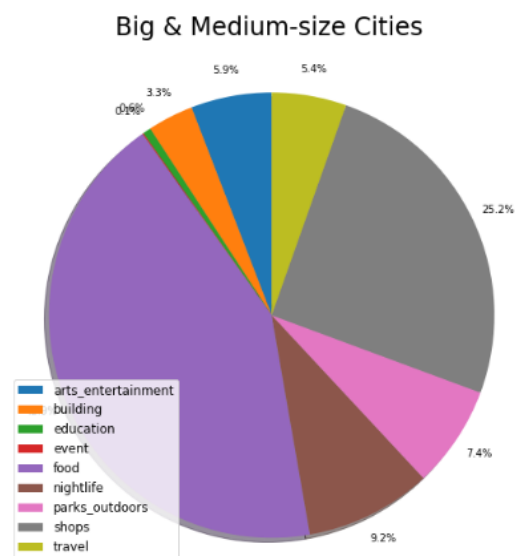
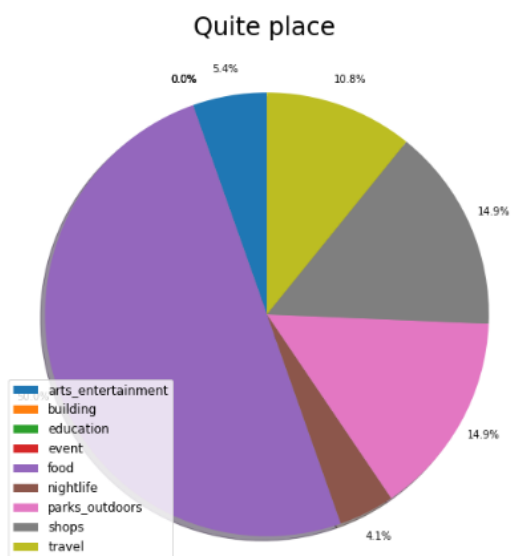
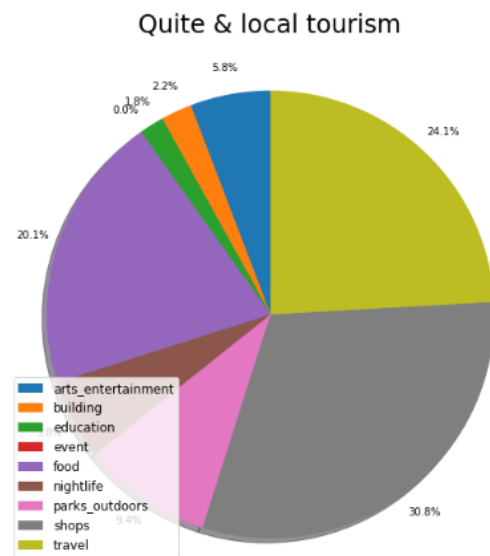
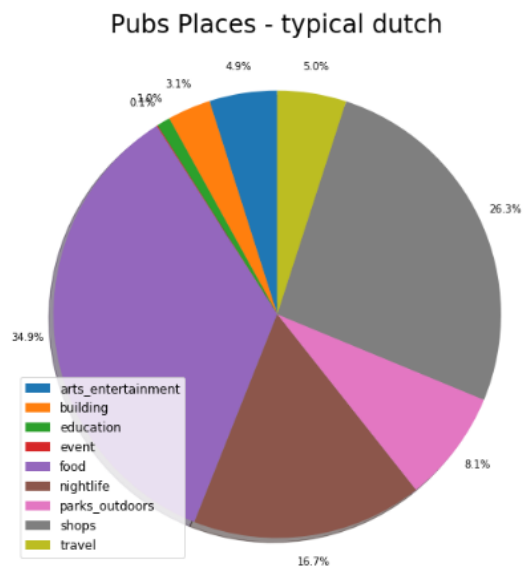
Those are the key observations that we can make based on this visualization:

- The big and medium sized cities are pretty similar between the north and the south w.r.t. popular venues (purple)
- For the smaller cities there is a dissimilarity between the north (orange) and the south (green).
- There are a small number of “special” cities (blue and red), that we will keep out of the discussions

We know now where the different kinds of cities are positioned on the map, but what we don't know yet is what makes those cities different w.r.t. popular venue types. Below you can find the what the most popular venue categories are for each cluster (city type). These pie charts show - for the 9 level-1 venue categories – what the most popular categories are for a given cluster. The level-1 venue categories can be found in de legends.

We took all the venues of all the cities belonging to a given cluster, calculated the total number of occurrences for each of the level-1 venue categories, and showed the percentage for each category on the pie chart.

Social and Cultural Differences between Regions in Belgium



Social and Cultural Differences between Regions in Belgium

The above already gives us a basic idea about what the most popular venues are for each cluster, but it only took into account the level-1 venue category. A deeper insight can be obtained by using all venue category levels. There are so many of them, which makes it impossible to visualize using a pie chart. Below you can find the top-10 most popular venue categories for each cluster.

Pubs Places - typical dutch	Quite & local tourism	Quite place	Big & Medium-size Cities	Food & Shops - typical french
Venue Category	Venue Category	Venue Category	Venue Category	Venue Category
Bar 722	Bus Stop 48	Bakery 18	Bar 470	Supermarket 272
Friterie 270	Pharmacy 11	Athletics & Sports 3	Supermarket 332	Italian Restaurant 68
Bakery 237	Friterie 11	Soccer Field 3	Italian Restaurant 318	French Restaurant 68
Supermarket 209	Athletics & Sports 10	Pharmacy 3	Bakery 303	Friterie 64
Bus Stop 145	Supermarket 8	Bookstore 3	Restaurant 302	Restaurant 62
Restaurant 120	Bakery 6	Friterie 3	Friterie 272	Bakery 57
Sandwich Place 113	Bar 6	Italian Restaurant 2	French Restaurant 235	Bar 56
Plaza 103	Park 6	Playground 2	Sandwich Place 232	Pizza Place 51
Pub 97	Sports Bar 4	Road 2	Plaza 178	Sandwich Place 35
Bistro 96	Restaurant 4	Park 2	Pizza Place 157	Fast Food Restaurant 30

Key observations:

- What stands out for the typical dutch cities is the popularity of the knight life (bars, pubs, friteries – where you can eat freanch-fries). Surprisingly, it is even larger than for the big cities.
- The french cities tend to be more quite when it come to nightlife.

7 Discussion

In Belgium the differences between the two communities is almost constantly a very hot topic (the only exception is when our national soccer team is playing – then we are still one country 😊). I was kind of sceptic about whether this would also be visible in the data – the facts.

To my own surprise, the data and the machine learning algorithms actually show some major differences.

- The popular venues in major cities in the north and the south are pretty similar.
- Apparently, the smaller Dutch-speaking cities more vivid, where nightlife is even more popular than in the bigger cities.
- In the smaller French-speaking cities people prefer going to a sports venue or having a quite evening in a restaurant.

I actually take the results with a grain of salt!

For the following reasons:

- There might not be is not sufficient data in the foursquare database, so that we can speak of hard evidence for proving the hypothesis.
- Due to different level of popularity of Foursquare for the two regions, there might be an imbalance between data available for the north and for the south.
- The fact that nightlife (pubs and bars) is more popular in the smaller Dutch-speaking cities than in the bigger cities might be explained by the fact that smaller cities have less other types of venues (e.g. cultural or historic venues)

Nevertheless, the visualised results clearly show a difference! The reason of the difference is less straightforward, however.

8 Conclusions

The hypothesis that we wanted to validate was:

There are important social and cultural differences between northern (Dutch-speaking) and southern (French-speaking) cities in Belgium.

We were able to draw the following conclusions from the data using machine learning techniques:

- For the bigger cities there are no visible differences between the north and the south w.r.t. popular venues
- For the smaller cities, however, there are dissimilarities between the north and the south.
 - What stands out for the typical dutch cities is the popularity of the knight life (bars, pubs, friteries – where you can eat freanch-fries). Surprisingly, it is even larger than for the big cities.
 - The french cities tend to be more quite when it come to nightlife.

-

Some advice for the companies located in Brussels that are working with a mix of employees from both parts of the country:

A Team building event like “let’s all go the pub this evening and quickly grab some french-fries on our way home” might be appreciated a lot by the Dutch-speaking employees, but might not work for the French-speaking employees. Maybe the latter prefer a cosy dinner in a good restaurant. [please don’t take this in the strict sense]