Social and Cultural Differences between Regions in Belgium

[Coursera – IBM Data Science – Capstone project]

Sven De Smit

innocom
INNOVATION MEETS COMMITMENT

Business Understanding

- Problem statement
  - Many Belgian 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 take into account cultural and social 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.

Analytical 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.

innocom
INNOVATION MEETS COMMITMENT

Data acquisition

- Belgian cities data [source: wikipedia https://nl.wikipedia.org/wiki/Tabel_van_Belgische_gemeenten]
  - A table that published on the Dutch Wikipedia that gives an overview of the cities in Belgium with some key metadata (name of the city, province, number of habitants, acreage, habitants per km², prosperity index
- Geo-location data [source: Python geopy package]
  - 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.
- Multi-language city names [source: economie.fgov.be KBO-codes-identificatie.xls]
  - 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.
- City venues data [source: Foursquare Places API]
  - A good source of information for venues all over the world is Foursquare. One can easily find popular venues in a city or a location by using the explore function in the Place API.
- Geo-location data for city borders in Belgium [source: municipalities-belgium.geojson]
  - 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.

Data cleansing and preparation

"Most popular venues for each city" dataset.
This dataset will be used by the clustering algorithm!

Cleansed Cities dataset

- Numbers reformatted
- Multi-language support added
- Geojson name created

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', '51.3778412', '4.2175769', '4.4...	Anvers	Antwerpen	Antwerpen	01.01.0001	31.12.9999	Antwerpen
1679	102.3	Oost-Vlaanderen	51.053829	3.725012	['50.9795422', '51.187948', '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', '50.6881961', '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	Schaarbeek	-	01.01.0001	31.12.9999	Schaarbeek#Schaerbeek
6747	63.8	Brussel	50.839098	4.329653	['50.8070598', '50.8505449', '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	Brügg			
631	99.9	Namen	50.466528	4.866189	['50.3872825', '50.5313007', '4.723053', '4.98...	Namur	Namen				
1794	113.6	Vlaams-Brabant	50.879202	4.701168	['50.8242096', '50.9440707', '4.640295', '4.77...	Louvain	Leuven	Löwe			

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

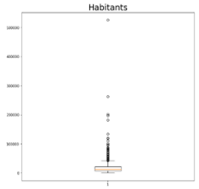
Enriched Venues dataset

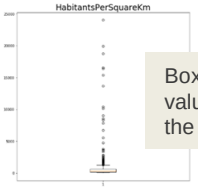
- Venue categories levels extracted

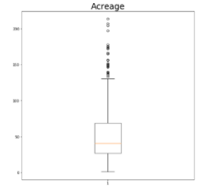
Out[20]:

	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/fi...	Fish & Chips Shop	food	food fishandchips_

Methodology – Insight & Visualization – Cities Dataset





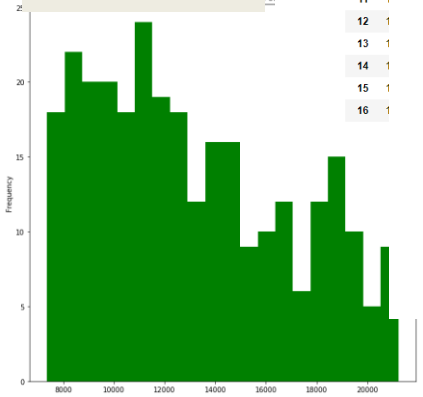


Box-plot showing values distribution for the key parameters

Top cities with most number of habitants

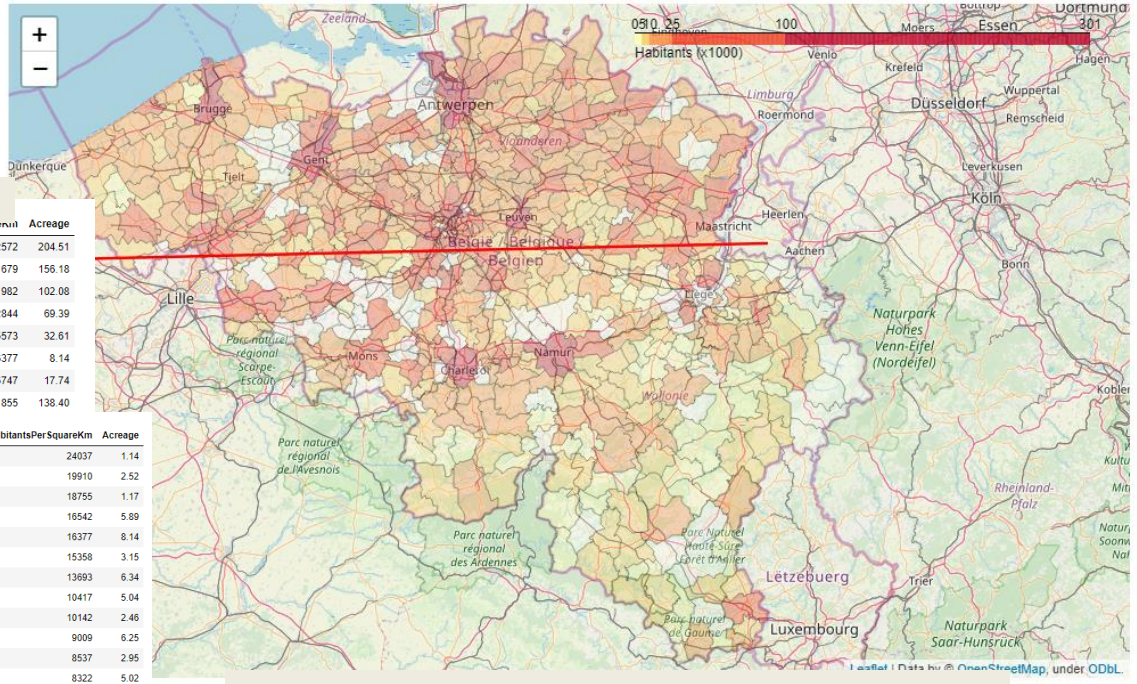
	ID	Name	Habitants	HabitantsPerSquareKm	Acreage
0	1	Antwerpen	525935	2572	204.51
1	2	Gent	262219	1679	156.18
2	3	Charleroi	202267	1982	102.08
3	4	Luik	197327	2844	69.39
4	5	Brussel	181726	5573	32.61
5	6	Schaarbeek	133309	16377	8.14
6	7	Anderlecht	119714	6747	17.74
7	8	Brugge	118325	855	138.40
8		Out[39]:			
9	1				
10	1				
11	1				
12	1				
13	1				
14	1				
15	1				
16	1				

Histogram habitants without outliers



	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-Woluwe	56660	7842	7.22
6	7	Anderlecht	119714	6747	17.74
4	5	Brussel	181726	5573	32.61
40	41	Sint-Pieters-Woluwe	41824	4725	8.85
64	65	Oudergem	34013	3765	9.03

Top cities with most number of habitants per km²



Map with showing number of habitants per city

Methodology – Insight & Visualization – 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' ... ]
```

```
----'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
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
```

innocom
INNOVATION MEETS COMMITMENT

Methodology – k-means clustering machine learning

A 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 dataset**.

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 (see next slide).

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

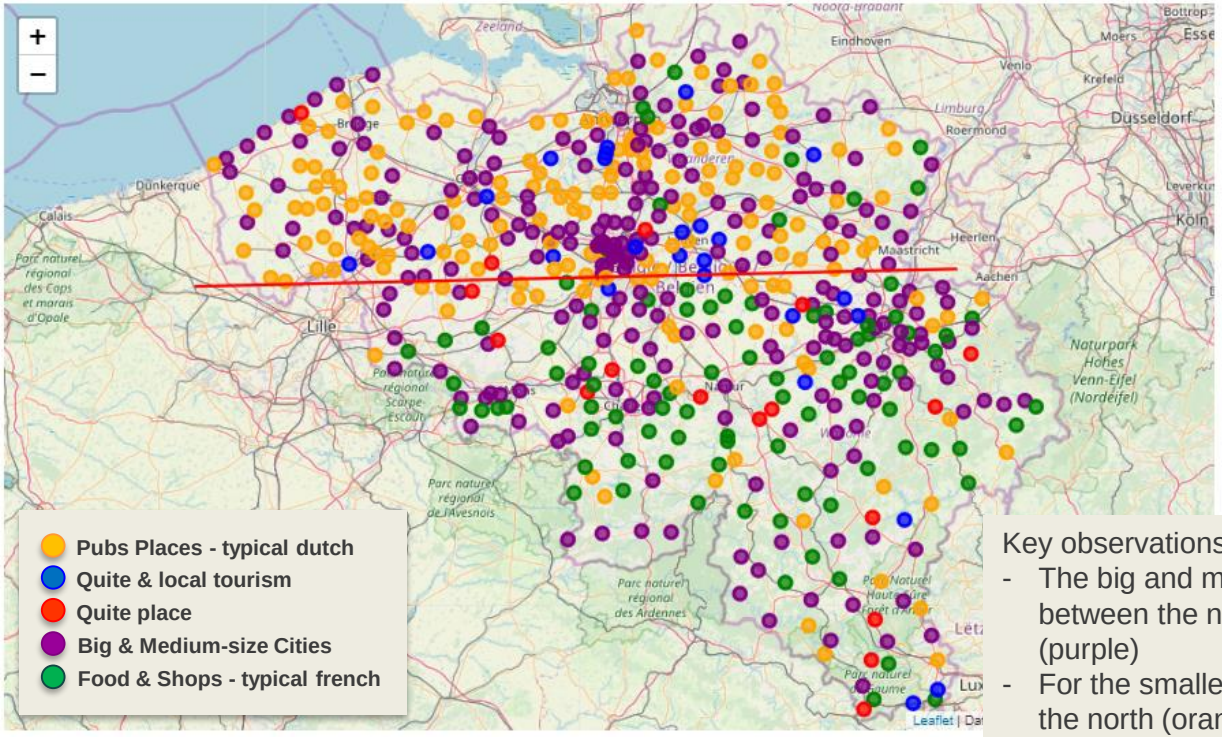
After experimentation we saw that using **5** as the number of **clusters**. The algorithm yields the following results:

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

	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

	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

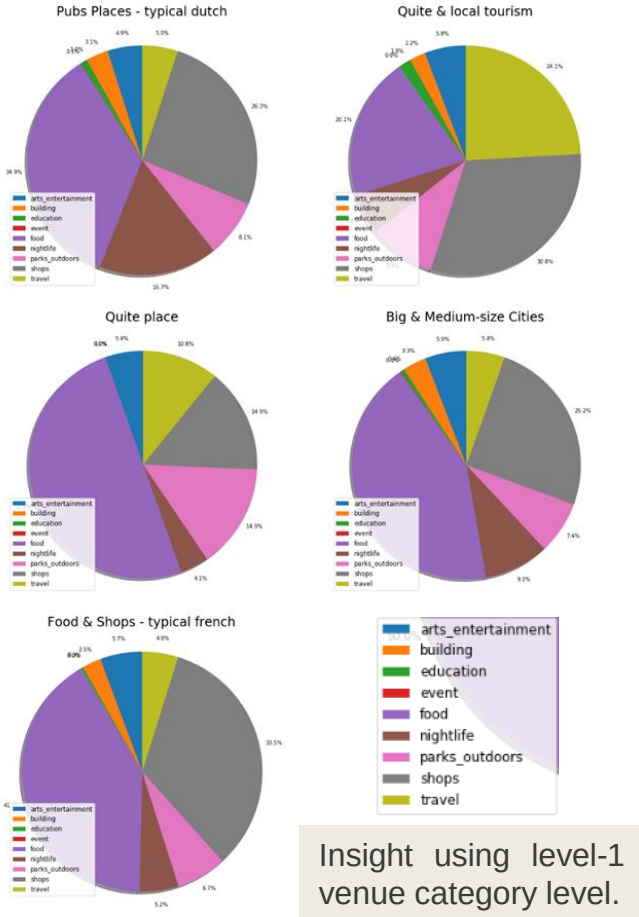
Results – City clusters on a map of Belgium



Key observations :

- 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

Results – Insight into the differences between clusters



Insight using level-1 venue category level.

A deeper insight using all venue category levels.

Pubs Places - typical dutch

Venue Category	
Bar	722
Friterie	270
Bakery	237
Supermarket	209
Bus Stop	145
Restaurant	120
Sandwich Place	113
Plaza	103
Pub	97
Bistro	96

Quite & local tourism

Venue Category	
Bus Stop	48
Pharmacy	11
Friterie	11
Athletics & Sports	10
Supermarket	8
Bakery	6
Bar	6
Park	6
Sports Bar	4
Restaurant	4

Quite place

Venue Category	
Bakery	18
Athletics & Sports	3
Soccer Field	3
Pharmacy	3
Bookstore	3
Friterie	3
Italian Restaurant	2
Playground	2
Road	2
Park	2

Big & Medium-size Cities

Venue Category	
Bar	470
Supermarket	332
Italian Restaurant	318
Bakery	303
Restaurant	302
Friterie	272
French Restaurant	235
Sandwich Place	232
Plaza	178
Pizza Place	157

Food & Shops - typical french

Venue Category	
Supermarket	272
Italian Restaurant	68
French Restaurant	68
Friterie	64
Restaurant	62
Bakery	57
Bar	56
Pizza Place	51
Sandwich Place	35
Fast Food Restaurant	30

Key observations:

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

innocom
INNOVATION MEETS COMMITMENT

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!
  - 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.

Conclusion

- 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 quicly eat some french-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]

innocom
INNOVATION MEETS COMMITMENT

The End

inno·com
INNOVATION MEETS COMMITMENT