Clustering Municipalities in Brussels

Matheus Mafra Oliveira Andrade

IBM Data Science Professional Certificate – Coursera – Oct/2020

1. INTRODUCTION

Belgium is the 13rd most populous country in Europe and the 11th most developed in the continent. It is also known by its linguistic and cultural diversity, and we cannot forget the Belgian tradition in producing handcrafted beer. Its capital, Brussels, is famous by the cafés, pubs, and bars, where the locals and tourists can spend time with friends and enjoy a good Belgian beer.

With this in mind, an entrepreneur has a dream of opening a new bar in Brussels where he can sell different beers, but he does not know where he should open it. His priority is to avoid strong competition, because he wants to gain trust from clients and a good market share before expanding his business. Thus, he wants to open the bar somewhere distant to other bars. Also, his idea is to serve mostly local people who wouldn't like to travel so far to enjoy a night out.

He is now analyzing the municipalities in the city's metropolitan area and, with the help of Foursquare dataset about places in Brussels, the entrepreneur will cluster the city's municipalities to check in which of them there are less bars in comparison to other types of venues, and, in other words, the best municipalities to open a new bar.

2. DATA DESCRIPTION

In order to complete the enterpreneur's study, he will need data from three different sources:

- A table containing the names of all Brussels' municipalities, based on data retrieved in 2015 by the Brussels Regional Informatics Centre (CIBG/CIRB). The table can be found on the Wikipedia page:

https://en.wikipedia.org/wiki/List_of_municipalities_of_the_Brussels-Capital_Region;

- Coordinates for each municipality, collected in the Python Geopy library;

- A table containing the names of venues in Brussels, along with their category, latitude, and longitude.

3. METHODOLOGY

The methodology section will be divided into four parts that represent the chronological order of the project. In each of them, processes, statistical analyses, and tools that have been used to successfully complete the work will be explained.

3.1. Data Cleaning – Municipalities Demographic Data:

After understanding the problem to be solved (which Brussels' municipality is the better place to open a bar), the first thing to do is collecting and cleaning demographic data about the 19 municipalities in the city. These data were collected in the Wikipedia page about Brussels municipalities and a Python library called BeautifulSoup was used to parse the HTML code and convert the table displayed on the website into a Pandas dataframe.

	Unnamed: 0	French name	Dutch name	Flag	CoA	postcode	Population(1/1/2017)	Area	Population density(km²)	Ref.
0	1	Anderlecht	Anderlecht	NaN	NaN	1070	118241	2 (6.8 sq mi)	6680	[7]
1	2	Auderghem	Oudergem	NaN	NaN	1160	33313	2 (3.5 sq mi)	3701	[8]
2	3	Berchem-Sainte-Agathe	Sint-Agatha-Berchem	NaN	NaN	1082	24701	2 (1.1 sq mi)	8518	[9]
3	4	Bruxelles-Ville*	Stad Brussel*	NaN	NaN	1000102011201130	176545	2 (12.6 sq mi)	5415	[10]
4	5	Etterbeek	Etterbeek	NaN	NaN	1040	47414	2 (1.2 sq mi)	15295	[11]

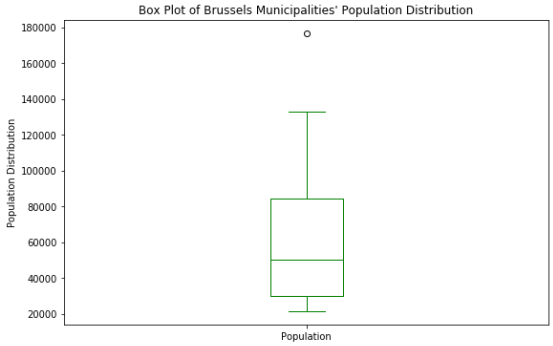
The image above shows the first five rows of the dataframe. It can be seen that there are some columns with duplicate, missing and useless information. So, for the analysis to be easier and more concise, those columns were dropped from the dataframe. More specifically, columns: "Unnamed: 0", "Dutch name", "Flag", "CoA", "postcode" and "Ref."

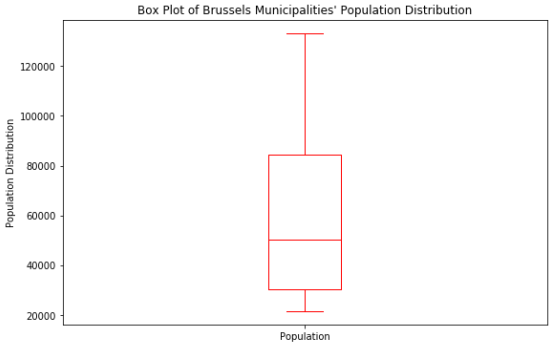
Besides that, since the column "Area" represents a continuous variable, parenthesis and units would jeopardize further calculus with this variable. Thus, those characters were deleted. Finally, Area and Population Density must be in the same measure, so the first one was converted to km^2 when multiplied by 2.58998811. The first five rows of the resulting dataframe, after some minor changes (like changing columns names) should look like this:

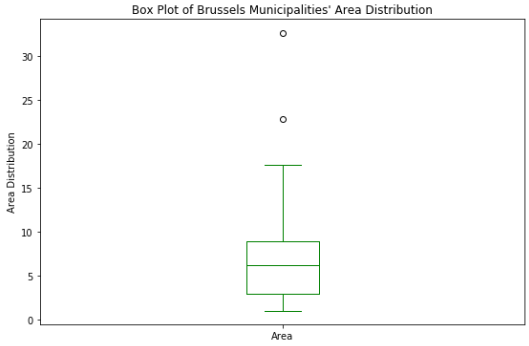
	Municipality	Population	Area	Pop density
0	Anderlecht	118241.0	17.611919	6713.691961
1	Auderghem	33313.0	9.064958	3674.920345
2	Berchem-Sainte-Agathe	24701.0	2.848987	8670.099472
3	Bruxelles-Ville	176545.0	32.633850	5409.873459
4	Etterbeek	47414.0	3.107986	15255.539790

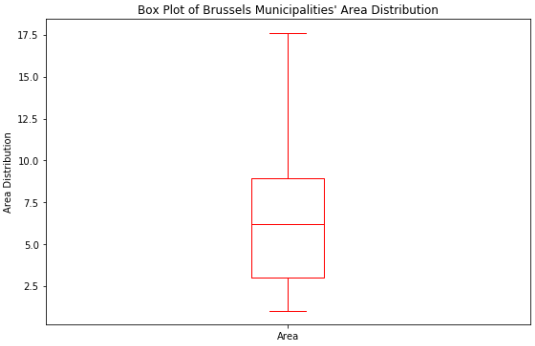
The last thing to do with this dataframe is a brief statistical description of the data:

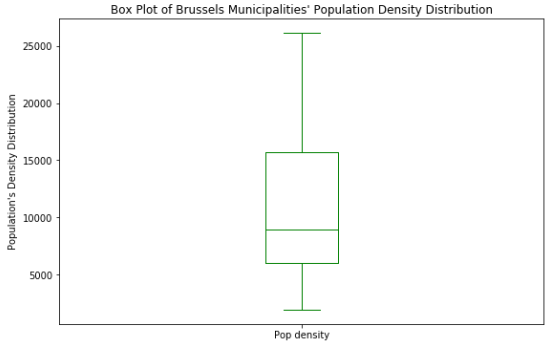
	Population	Area	Pop density
count	19.000000	19.000000	19.000000
mean	62716.000000	8.465172	10725.764897
std	42681.784033	8.071577	6365.413739
min	21609.000000	1.035995	1920.549357
25%	30214.000000	2.978486	6061.782710
50%	50471.000000	6.215971	8968.187889
75%	84275.500000	8.935459	15738.349348
max	176545.000000	32.633850	26172.900076

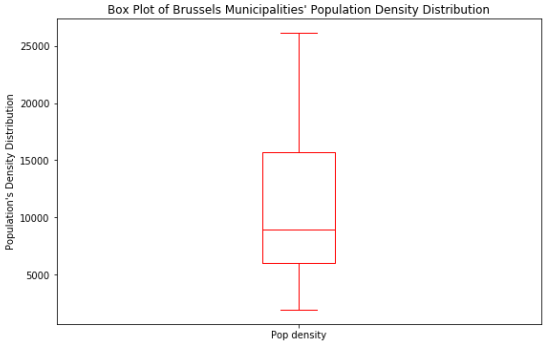












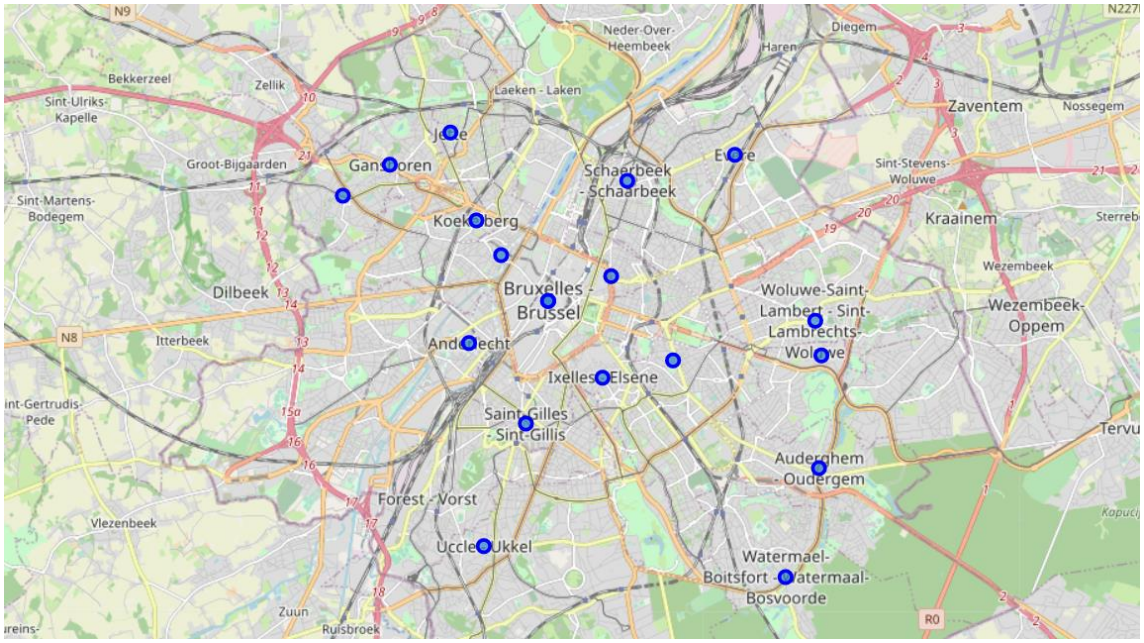
Obs: Box plots in green shows the distribution of the data with outliers, in red outliers are not considered.

3.2. Data Collecting – Municipalities Coordinates:

The next step is a necessary transition between gathering demographic data and venues data: getting geographical coordinates. To do that, Python package Geopy was used and the coordinates for each municipality were joined into the Pandas dataframe with demographic data. Its first five rows should look like this:

	Municipality	Population	Area	Pop density	Latitude	Longitude
0	Anderlecht	118241.0	17.611919	6713.691961	50.839098	4.329653
1	Auderghem	33313.0	9.064958	3674.920345	50.817236	4.426898
2	Berchem-Sainte-Agathe	24701.0	2.848987	8670.099472	50.864923	4.294673
3	Bruxelles-Ville	176545.0	32.633850	5409.873459	50.846557	4.351697
4	Etterbeek	47414.0	3.107986	15255.539790	50.836145	4.386174

After that, to better visualize the spatial distribution of Brussels' municipalities, a map was created with the Python package Folium, and was based on the coordinates displayed on the dataframe:



3.3. Data Analysis – Venue Data:

The last thing to do before building a model is getting data from different types of venues in Brussels. Data represents venues in a 750 meters radius. They were collected from the Foursquare API and information about venue name, category and where it is located were stored into another Pandas dataframe:

	Municipality	Municipality_Latitude	Municipality_Longitude	Venue	Venue_Latitude	Venue_Longitude	Venue_Category
0	Anderlecht	50.839098	4.329653	Chez Rosario	50.836240	4.331007	Deli / Bodega
1	Anderlecht	50.839098	4.329653	Crep' & Cream	50.839050	4.330786	Creperie
2	Anderlecht	50.839098	4.329653	Brasserie Cantillon Brouwerij (Cantillon - Bro...	50.841487	4.335451	Brewery
3	Anderlecht	50.839098	4.329653	Maharaja Tandoori Restaurant I	50.839015	4.332212	Indian Restaurant
4	Anderlecht	50.839098	4.329653	Boeremet	50.842882	4.326992	Cocktail Bar

After checking the number of venues in which municipality, a new dataframe was created using one hot encoding. It displays the same number of rows (each one representing a venue), and each column represents one venue category. The data shown can be either 0 or 1. If a cell is marked as 1, it means that the venue (row) belongs to the category in the column where the cell is marked as 1. To illustrate the situation better, the image shows the dataframe:

	Municipality	African Restaurant	American Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auto Garage	...	Vegetarian / Vegan Restaurant	Video Game Store	Video Store	Vietnamese Restaurant	Volley C
0	Anderlecht	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
1	Anderlecht	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
2	Anderlecht	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
3	Anderlecht	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
4	Anderlecht	0	0	0	0	0	0	0	0	0	...	0	0	0	0	

The venues shown in the image above do not belong to the categories displayed, since they are all marked as 0.

After that, a new dataframe was created. It shows the average frequency of each venue category by municipality.

	Municipality	African Restaurant	American Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auto Garage	...	Vegetarian / Vegan Restaurant	Video Game Store	Video Store
0	Anderlecht	0.00	0.00	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	...	0.023256	0.000000	0.000000
1	Auderghem	0.00	0.00	0.00	0.000000	0.000000	0.000000	0.000000	0.015385	0.00000	...	0.000000	0.000000	0.000000
2	Berchem-Sainte-Agathe	0.00	0.00	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	...	0.000000	0.000000	0.000000
3	Bruxelles-Ville	0.00	0.01	0.00	0.000000	0.000000	0.000000	0.010000	0.000000	0.00000	...	0.000000	0.000000	0.000000
4	Etterbeek	0.00	0.00	0.01	0.000000	0.010000	0.000000	0.010000	0.000000	0.00000	...	0.000000	0.000000	0.000000

Finally, total frequency of each category was ranked by the five most common ones by municipality. For example, these are the five most common types of venues in Anderlecht:

```
----Anderlecht----
                venue  freq
0               Hotel  0.10
1         Coffee Shop  0.08
2      Sandwich Place  0.07
3  French Restaurant  0.05
4       Train Station  0.03
```

3.4. Running K-Means Clustering:

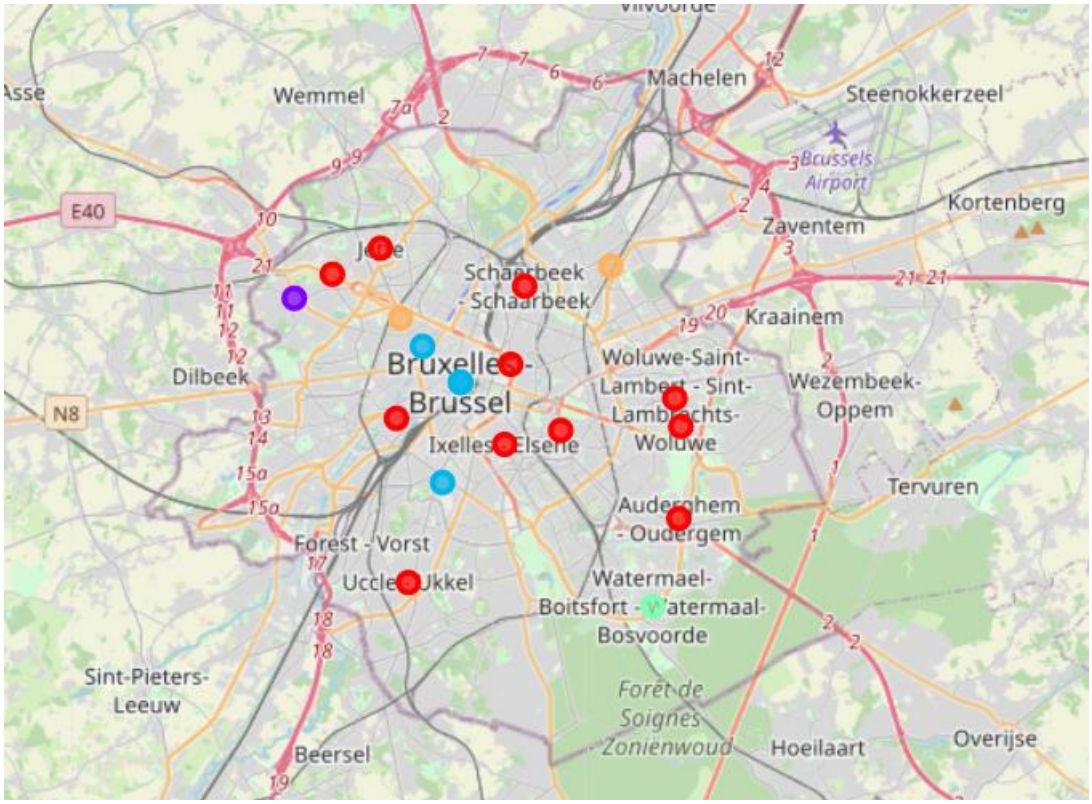
After a long process of data cleaning, collecting and analysis, the machine learning model can be finally built. Since the entrepreneur wants to know a group of municipalities with similar characteristics that represents a good place to open a bar, a clustering algorithm, like K-Means, should be used. Also, since the information needed is not displayed on any dataframes, the model shall learn from the data to create the groups of municipalities. Therefore, an unsupervised model, like K-Means Clustering seems to be a good one.

The municipalities were divided into 5 different clusters. To select the most similar municipalities to put inside the same cluster, and, consequently, the most distinct to put in different clusters, the algorithm considered the venues that appeared in each one's top 5 ranking. For example, municipalities with bars appearing in the ranking should be put in the same cluster; but, if in one of them there are lots of bars, and in another one there are lots of schools, they should be put in different clusters.

The model was then fitted, and it created 5 cluster labels (from 0 to 4). After that, it assigned a label for each municipality, and this information was stored into a Pandas dataframe, along with the most common venues in each municipality:

	Municipality	Population	Area	Pop density	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Anderlecht	118241.0	17.611919	6713.691961	50.839098	4.329653	0	Hotel	Coffee Shop	Sandwich Place	French Restaurant	Bar
1	Auderghem	33313.0	9.064958	3674.920345	50.817236	4.426898	0	Italian Restaurant	Fast Food Restaurant	Pizza Place	French Restaurant	Bakery
2	Berchem-Sainte-Agathe	24701.0	2.848987	8670.099472	50.864923	4.294673	1	Greek Restaurant	Tram Station	Gym	Restaurant	Bar
3	Bruxelles-Ville	176545.0	32.633850	5409.873459	50.846557	4.351697	2	Bar	Beer Bar	Chocolate Shop	Plaza	Thai Restaurant
4	Etterbeek	47414.0	3.107986	15255.539790	50.836145	4.386174	0	Italian Restaurant	Bakery	Restaurant	Plaza	Wine Bar

Next, to visualize the clusters better, a new map was created using Folium, however, this time, each municipality was displayed in a different color, corresponding to its cluster label:



Red: Cluster 0 Violet: Cluster 1 Cyan: Cluster 2 Green: Cluster 3 Orange: Cluster 4

4. RESULTS

Finally, after each municipality was assigned to a cluster, each of the five groups should be analyzed separately. For each cluster, a Pandas dataframe was created to show the top five venue categories and their frequency:

Cluster 0:

	Sum	Frequency
Plaza	8.0	0.145455
Italian Restaurant	6.0	0.109091
Bakery	5.0	0.090909
Supermarket	4.0	0.072727
Bar	3.0	0.054545

Cluster 1:

	Sum	Frequency
Tram Station	1.0	0.2
Restaurant	1.0	0.2
Gym	1.0	0.2
Greek Restaurant	1.0	0.2
Bar	1.0	0.2

Cluster 2:

	Sum	Frequency
Bar	4.0	0.20
Plaza	3.0	0.15
Thai Restaurant	2.0	0.10
Chocolate Shop	2.0	0.10
Beer Bar	2.0	0.10

Cluster 3:

	Sum	Frequency
Restaurant	1.0	0.2
Park	1.0	0.2
Italian Restaurant	1.0	0.2
Ice Cream Shop	1.0	0.2
Chinese Restaurant	1.0	0.2

Cluster 4:

	Sum	Frequency
Supermarket	2.0	0.2
Snack Place	2.0	0.2
Sandwich Place	1.0	0.1
Park	1.0	0.1
Hotel	1.0	0.1

5. DISCUSSION

However, the idea in this project is to choose better places to open a bar based mostly on possible competition. Therefore, places in which there are less bars or other types of venues to enjoy the nightlife are good choices for the entrepreneur. Going back to the Results session, it can be seen that bars appear in the top five rankings for clusters 0, 1, and 2. Thus, the entrepreneur should select a municipality located on the cluster 3 or 4.

Making a deeper analysis, in cluster 3 there are lots of restaurants, mostly Italian and Chinese. Despite restaurants not representing direct competitors for bars, they can capture some customers who want to go out for dinner instead of having a drink at a good bar. So, places with lots of restaurants, like municipalities in cluster 3 are not ideal places to open a bar.

Finally, taking a better look in cluster 4, there are no bars and/or restaurants on the top 5 venue ranking. Thus, the entrepreneur should choose between the municipalities that are grouped into this cluster: Evere and Koekelberg. To make a more accurate decision, both municipalities should be analyzed:

	Municipality	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
5	Evere	4	Supermarket	Snack Place	Hotel	Brasserie	Sandwich Place
10	Koekelberg	4	Supermarket	Bar	Park	Gym	Snack Place

Since the 2nd most common venues in Koekelberg are bars, Evere might be the better place for the entrepreneur to open his business.

6. CONCLUSION

Of course, some points on the project and some assumptions made are not perfect. Information about average income, age, and interest in bars of each municipality's customers should be used to make a better decision about a good place to open this type of venue.

However, it can be used as a good basis for decision making, since the clustering algorithm grouped different municipalities where the entrepreneur should have weaker competition. Besides that, this project can serve as a starting point for other studies regarding similar problems that require a more complex analysis, model evaluation and product building.