Matheus Mafra
Oliveira Andrade

IBM Data Science
Professional Certificate

Coursera – Oct/2020

CLUSTERING MUNICIPALITIES IN BRUSSELS

1. Business Problem

- Brussels enterpreneur wants to open a bar;
- He wants to open it somewhere with less competition. In other words, somewhere distant from other bars and/or restaurants.
- To do that, he'll use a clustering machine learning algorithm to choose a municipality in Brussels which has a lack of bars and restaurants.

2. Data Description

- Table containing demographic data about Brussels municipalities;
  - Found in https://en.wikipedia.org/wiki/List_of_municipalities_of_the_Brussels-Capital_Region
- Geographic coordinates of each Brussels' municipality;
  - Gathered in Python Geopy library
- Table containing information about diferente venues in Brussels.
  - Found in Foursquare API

3. Methodology

- 3.1. Data Cleaning - Municipalities demographic data:

  - Parsing HTML code to collect data from Wikipedia page;
  - Removing unnecesary columns;
  - Changing “Area” data type from string to float;
  - Converting “Area” values from square miles to square kilometers;
  - Statistical description of the data.

3. Methodology

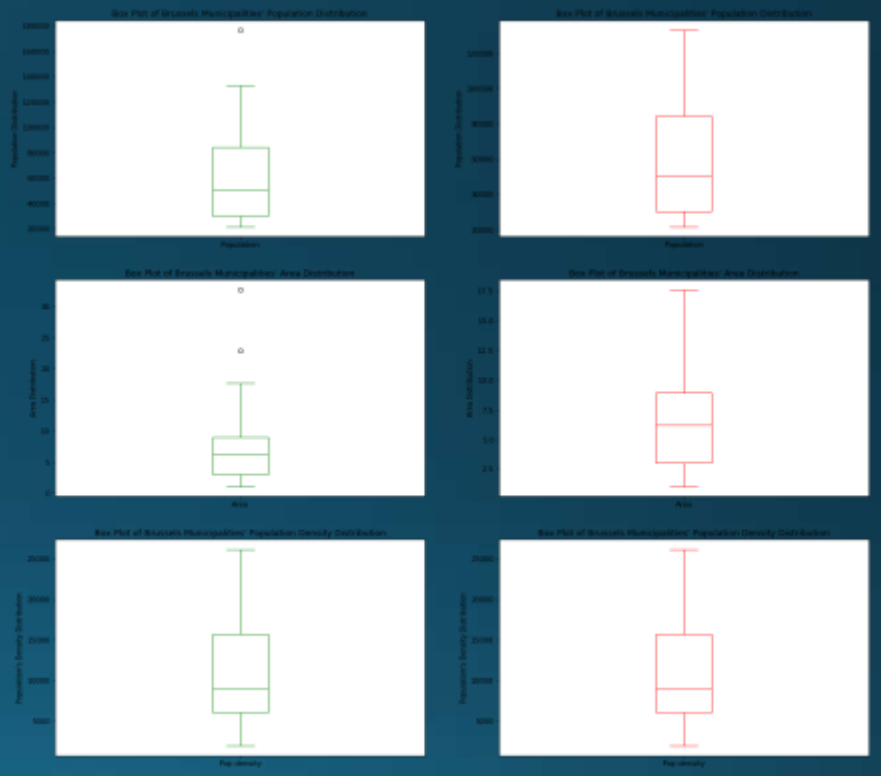
3.1. Data Cleaning - Municipalities demographic data:

	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

Final dataframe (after cleaning)

	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

Descriptive statistics



Box plots of population, area and population density, respectively
Green: with ouliers. Red: without outliers

3. Methodology

- 3.2. Data Collecting - Municipalities coordinates:

  - Joining coordinates for each municipality into the dataframe;
  - Visualizing municipalities distribution in a Folium map.

	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

3. Methodology

- 3.3. Data Analysis – Venue data:

  - Getting data about Brussels venues using Foursquare API;
  - Applying one hot encoding to determine each venue category;
  - Calculating the average frequency of each venue category by municipality;
  - Ranking top five venue categories in frequency by municipality.

3. Methodology

3.3. Data Analysis – Venue data:

	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	4.330798	Creperie
2	Anderlecht	50.839098	4.329653	Brassene 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	4.328992	Cocktail Bar

Venue data

	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

Average frequency 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	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	

One hot encoding

```
----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
```

Top 5 venue categories in Anderlecht

3. Methodology

- 3.4. Running K-Means Clustering:

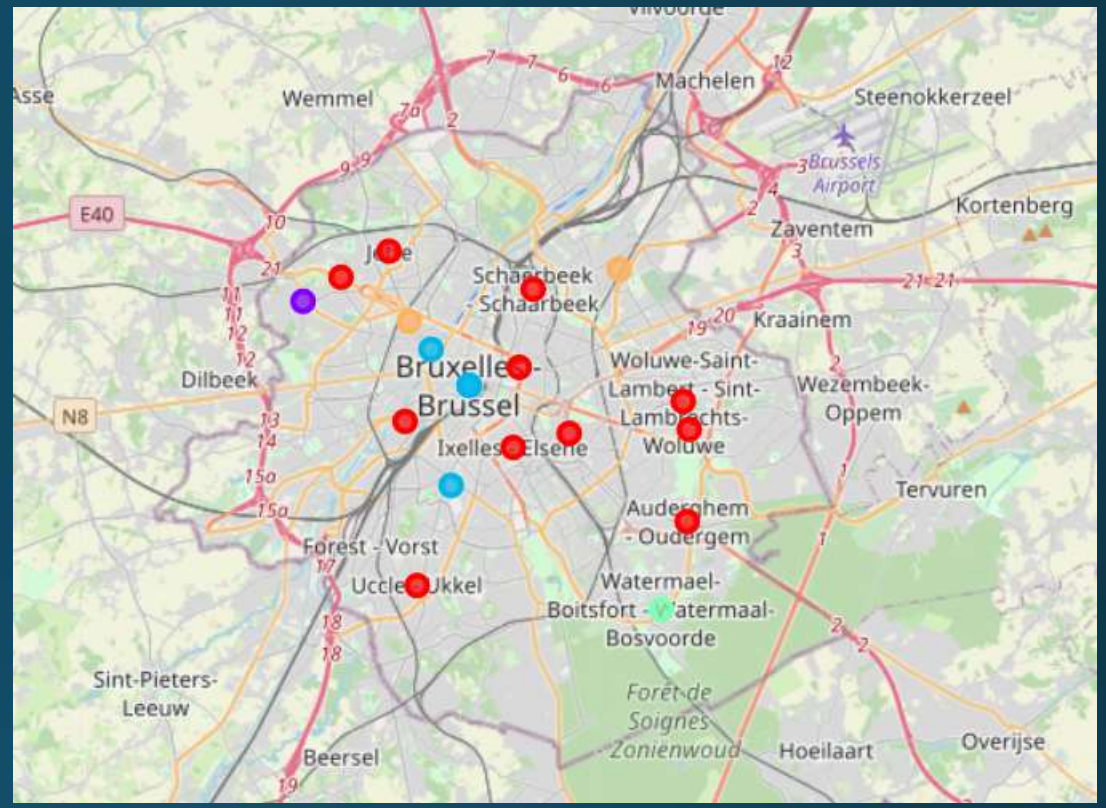
  - Objective: group municipalities with similar categories, distinct from other municipalities;
  - Fitting the model: the algorithm creates five diferente cluster labels;
  - Assigning each municipality into a cluster label;
  - Inserting the labels into the dataframe with the most common venues in each municipality;
  - Visualizing the municipalities in different clusters in a Brussels map.

3. Methodology

3.4. Running K-Means Clustering:

	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

Dataframe with cluster labels and most common venues



Map of Brussels

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

4. Results

- Analyzing each cluster separatelly:
  - Creating a ranking of the 5 most common categories in each one;
  - Selecting the ones in which there are no bar in the ranking (clusters 3 and 4).

	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 0

	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 1

	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 2

	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 3

	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

Cluster 4

5. Discussion

- Clusters 3 and 4 appear to have less competition;
- However, there are many restaurants in Cluster 3, what can represent na indirect type of competition;
  - People might want to go out to dinner instead of going to a bar.
- Since there are no bars and restaurants in Cluster 4 ranking, it seems to be the best option for the enterpreneur.

5. Discussion

- There are 2 municipalities in Cluster 4: Evere and Koekelberg;
- Bars appear as the 2nd most common venue in Koekelberg;

	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

- Therefore, the enterpreneur should start his business in Evere!

6. Conclusion

- Some assumptions and analysis might not be perfect;
- The criteria used to open a bar is really simplified;
  - Other factors, like income, age and interest have a strong impact in business decisions
- However, the study can serve as a basis for projects that require more complex analysis regarding similar problems.