The Battle of Neighborhoods

This notebook will be used for the coursera capstone project

A. Introduction

A.1. Description & Disscusion of the Background

There are about 136 cities in Belgium with at least 15000 inhabitants housing about half of the populiation. For someone who is deciding to move to Belgium they may want to understand the different options where they could decide to live and how these cities compare to one another.

In this project we will consider the following parameters:

- Price of Real Estate
- Size of the city
- Types of social activities

Other people that might benefit from this type of analysis could be locals moving to another city for work or family reasons who would like to live in a similar place as what they are used to and want to optimize the investment cost in the real estate they are buying or people that would like to open a new business to understand better what the competition looks like.

We will first investigate the city clustering based on the type of venues and secondly we will investigate clustering based on real estate and city size using different visuals

A.2. Data Description

To solve the problem we will use following datasets

1) *A list of all Belgian cities from Wikipedia.* We will extract all the cities and restrict the analysis to cities with at least 15000 inhabitants. One complexity to consider to link the data to different datasets will be that cities can have names in Dutch, French or both. This will need to be handled in the data cleansing. [1]

2) *A list of all geolocations* of the centre of the different cities and the postal codes. This dataset covers all locations, but we will only extract those places that have >15000 inhabitants based on dataset 1 [2]

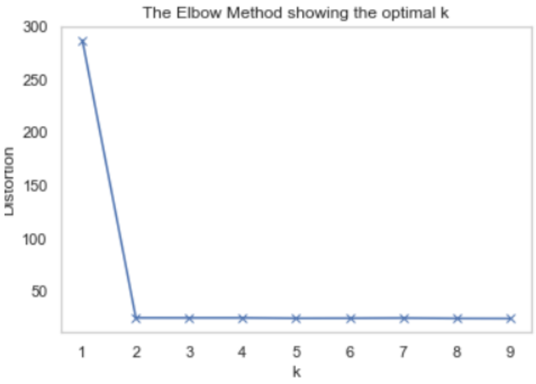
3) *An overview of housing prices*. This dataset from the Belgian Statistical Agency is quite elaborate and holds data from 2010 to 2017 and has a higher granularity than cities. For this exercise we will take the average housing prices of the most recent reported year aggregated per city.

B. Methodology

We will do the analysis in 2 steps

B.1 - Location analysis

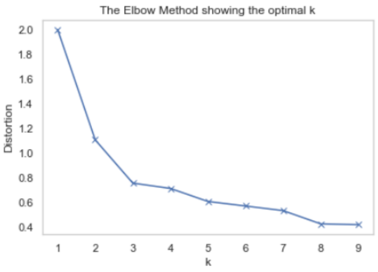
- Extract the data from datasets 1 & 2 to obtain a clean list of the different Belgian cities and their geolocations
    - We extract from dataset 1 both the Dutch and the French names along with the Province and the numer of Inhabitants. This gives us 136 different cities
    - We take a subset of the cities that have at least 15000 inhabitants, leaving 94 cities in the dataset and representing a total of 4.5 million inhabitants (this is about 45% of the total population)
    - Next we bring in dataset 2 to map the cities to their geolocation. The dataset 2 mixes French and Dutch citynames, so it was important to check for both languages if the location existed and add the corresponding gelocation. After this step 90 cities were left in the dataset. The remaining 4 where no geolocation was available, were all smaller cities where the name was slightly different. They were no further considered.
- Use the Foursquare API to gather data about the venues in a radius of 500m to each city centre
- Pre-process the data to map the cities to the presence of different venues
- Apply k-means clustering to the list of venues.
    - Based on the elbow method the best result would be expected for a k of 2.
    - A test was done at k=4 to assess if there could be any more details discovered based on the venues



- Visualize the clustered data on a map

B.2 - Price & Size analysis

- Add the real estate data to datasets 1 & 2
- Clean the dataset
- Apply k-means clustering to the inhabitants and housing price data. Optimal k was found to be 3.



- Visualize the clustered data on a map and in a scatter plat

Finally we will discuss the results and provide some guidance on the interpretation

C. Results & discussion

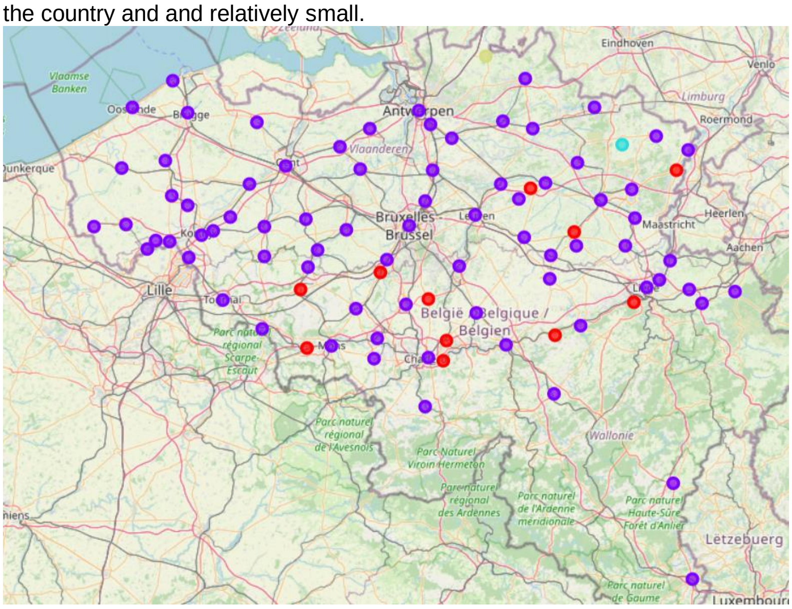
C.1 - Location analysis

Following plot shows the geolocation of the k-means clustering of the 90 Belgian cities and the colour corresponds to the category.

As the elbow method has shown there are really only 2 meaningful categories (the 3rd and 4th categorie only have one datapoint)

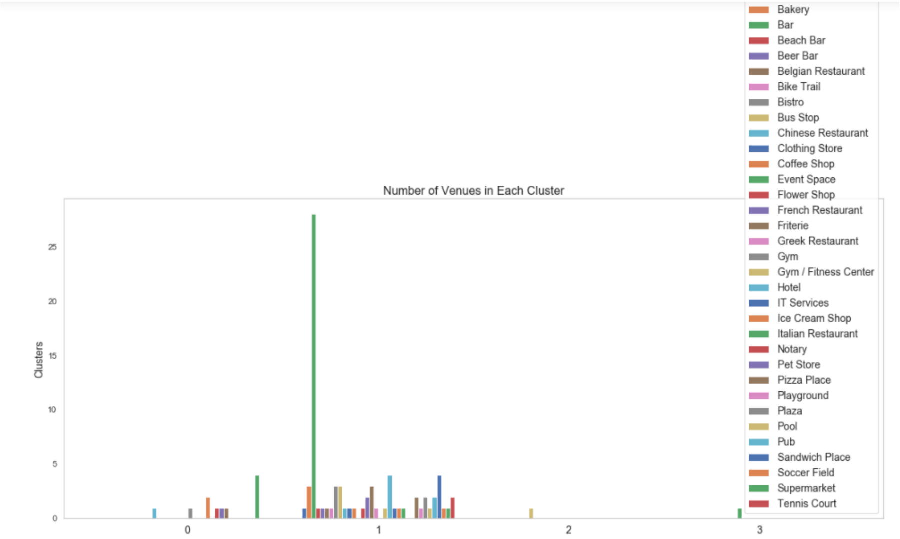
We find that almost all cities fall under category 2 (indicated by the purple colour)

There is a subgroup of cities that are in category 1 (indicated by red colour). Most of those are in the southern part of the country and and relatively small.


Zeeland
Eindhoven
Venlo
Vlaamse Banken
Oostende
Brugge
Antwerpen
Limburg
Roermond
Vlaanderen
Dunkerque
Gent
Bruxelles
Brussel
Leuven
Heerlen
Maastricht
Aachen
Kortrijk
Lille
Tournai
Liège
België Belgique / Belgien
Parc naturel régional Scarpe Escaut
Mons
Charleroi
Parc naturel régional de l'Avesnois
Wallonie
Parc Naturel Viroin Hermeton
Parc naturel régional des Ardennes
Parc naturel de l'Ardenne méridionale
Parc naturel Haute-Sûre Forêt d'Anlier
Letzebuerg
Parc naturel de Gaume
Luxembourg

Looking at the bar graph of the most frequent occurring venues for each cities we can mainly make 1 conclusion : the similarity of the city centres is mainly driven by the presence of bars. (Credits to sercan-yıldız for the idea and code of the graph)

This is no surprise, as Belgium has about 304 different active breweries.



The venues data from Foursquare applied to this dataset does not give a lot of insight. Given that some cities are quite small there is also a limited amount of data available to draw meaningful conclusion.

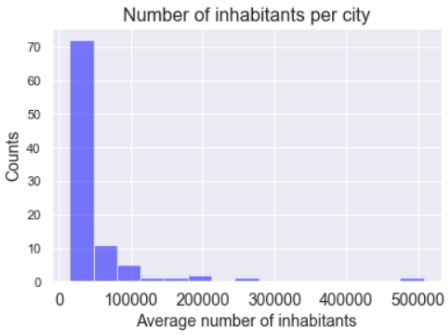
In the next section we will look at the clustering based on other features in the dataset.

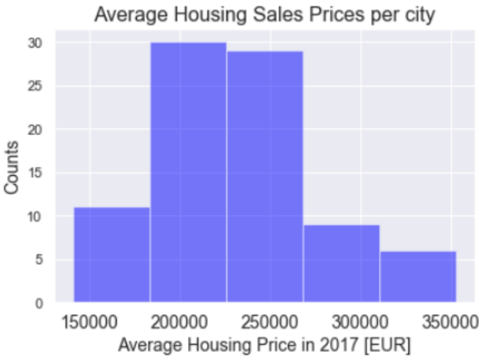
C.2 – Price & Size analysis

Below the graphs are shown of the amount of inhabitants per city grouped and the average housing prices.

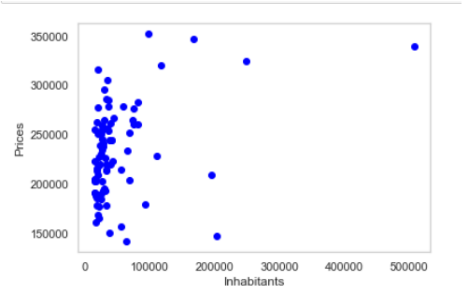
We can see that most cities are relatively small and only a handful of larger cities are within the dataset.

The housing prices are more evenly distributed and have more variation.



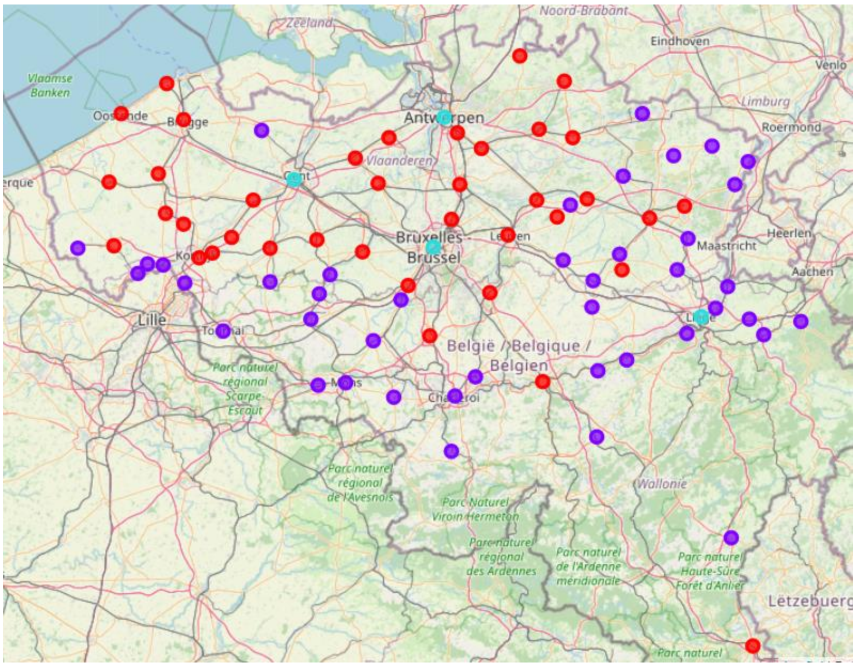


Putting the #inhabitants vs average housing price in a scatter plot indicates that there is some positive correlation between both parameters.

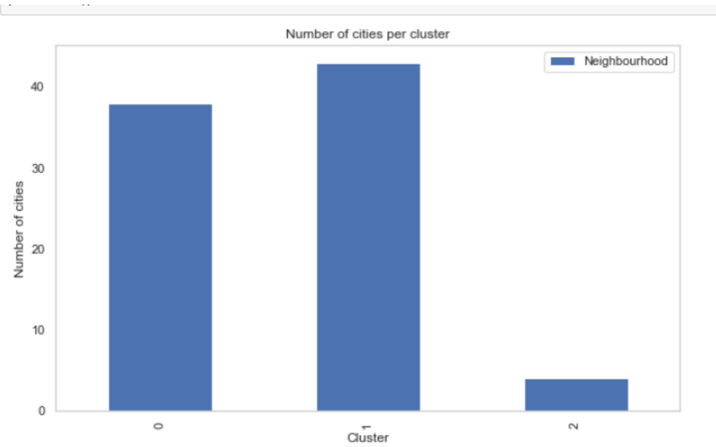


The k-means clustering of the different cities on a map clearly shows 3 main groups

- Cluster 0 (red) – we can clearly see these cities are mainly in the north of the country
- Cluster 1 (purple) – these cities are more in the south of the country
- Cluster 2 (cyan) – these are the more known cities in Belgium : Brussels / Antwerp / Gent / Liege

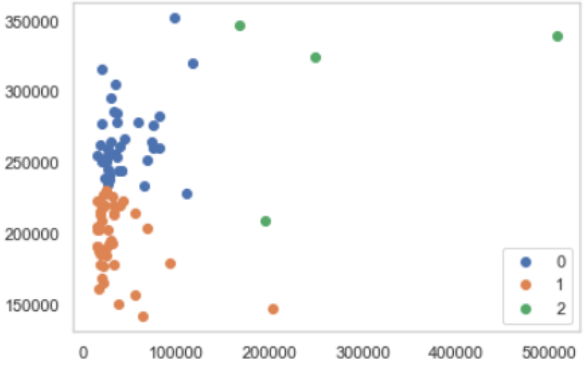

Noord-Brabant
Zeeland
Eindhoven
Venlo
Vlaamse Banken
Limburg
Roermond
Oostende
Brugge
Antwerpen
Vlaanderen
Gent
erque
Bruxelles - Brussel
Leuven
Heerlen
Maastricht
Aachen
Kortrijk
Lille
Tournai
Liège
België / Belgique / Belgien
Parc naturel régional Scarpe Escaut
Mons
Charleroi
Parc naturel régional de l'Avesnois
Parc Naturel Viroin-Hermeton
Parc naturel régional des Ardennes
Parc naturel de l'Ardenne méridionale
Wallonie
Parc naturel Haute-Sûre Forêt d'Anlier
Lëtzebuerg
Parc naturel

Below graph shows the number of cities in each category.



Plotting the scatterplot again with the number of inhabitants vs the average housing price with the cluster labels to define the colours shows us clearly that :

- Cluster 0 (blue) – these were the cities in the north of the contry. They tend to be relatively more expensive and have slightly larger populations
- Cluster 1 (orange) – these were the cities more south of the country. They tend to be relatively cheaper and have slightly smaller populations
- Cluster 2 (green) – these are the more known cities in Belgium : Brussels / Antwerp / Gent / Liege. These are much more populous and are most expensive.



D. Conclusion

From this analysis it can be concluded that in Belgium there are mainly 3 types of cities

- The north of the country is more populous and has more expensive housing
- The south of the country is less populous and has relatively cheaper housing prices
- There are a limited number of larger cities in terms of population that are most expensive. These are the more known tourist locations in Belgium

One thing is sure, no matter where you go. A bar will always be near.

E. Sources

Datasets:
[1] https://nl.wikipedia.org/wiki/Lijst_van_steden_in_Belgi%C3%AB
[2] https://github.com/jief/zipcode-belgium
[3] https://statbel.fgov.be/nl/open-data/verkoop-van-onroerende-goederen-gemeente-volgens-aard-op-het-kadastrale-plan-2010-2017