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.

We will do the analysis in 2 steps

step 1 - Location analysis

- Extract the data from datasets 1 & 2 to obtain a clean list of the different Belgian cities and their geolocations
- Use the foursquare API to gather data about the venues in a radius of 500m to each city centre
- preprocess the data to map the cities to the presence of different venues
- Apply k-means clustering to the list of venues
- visualize the clustered data on a map

step 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
- 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

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