

Battle of the neighborhoods

CAPSTONE Project

Marc De Decker | Coursera Data Science Professional | January 23, 2019

# Introduction

The audience for this project is an investor from Antwerp looking to open 3 new breakfast or coffee shops in the province of Antwerp, Belgium. The goal is to start small and expand in the future. That's why it's very important that they make the right choices for their initial investments. They want to start and open new shops in the province of Antwerp (Belgium) and depending on the success they want to open places in the rest of Flanders (Vlaanderen). The investor is not interested in doing business in Brussels or Wallonie which are the 2 other main parts of Belgium. The idea is to start the new coffee and/or breakfast shops in 3 dissimilar area's in respect to the number of similar existing shops but all being close to a shopping area, public transportation spots or other (touristic) attraction. It is important for the investor to understand if the success will depend on the number of shops that can be found in a radius of 1500m around the town center.

# Data

For this project, data obtained from a couple of websites are used. This includes a list of towns in Belgium and their corresponding Latitude and Longitude coordinates. These will be used as input to the Foursquare APIs to collect a list of Venues for the towns. The data will then be processed, analyzed and a recommendation will be made. Here is a list of websites leveraged for data collection:

<https://www.aggdata.com>

<https://api.foursquare.com/v2/venues/explore>

A list of all cities/villages that exist in the province of "Antwerpen" needed to be found along with the gps coordinates (latitude, longitude) of its town centers. A free list which contains this data could be found on the site of AggData (www.aggdata.com), a firm specialized in providing locational data. The data could be downloaded as a csv file and contains below info:

Postal Code: postal code of the town

Place Name: the name of the town

State: indicates where the town is located in Belgium (possible values are Bruxelles-Capitale, Vlaanderen and Wallonie)

State Abbreviation: abbreviation of the state being BRU, VLG or WAL

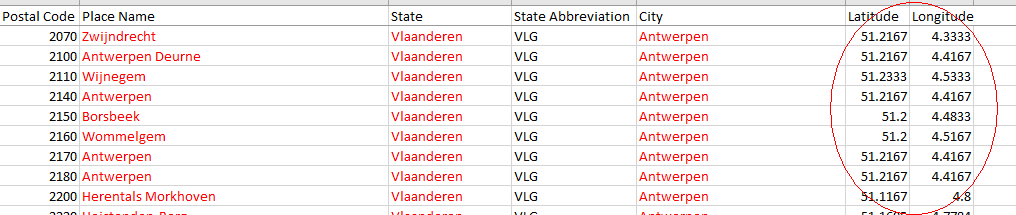
City: indicates in which Province of Belgium the town is located

Latitude: latitude gps coordinate of the town

Longitude: longitude gps coordinate of the town

***Note****: Belgium has 3 main parts (Bruxelles-Capitale, Vlaanderen and Wallonie) but the whole of Belgium is divided into 9 provinces. For this project the investor is only interested in the province of "Antwerpen", part of Vlaanderen. Vlaanderen is the dutch-speaking part of Belgium while Wallonie is the french-speaking part of Belgium. Bruxelles-Capitale is the capital of Belgium, where both dutch and french are spoken.*

The below shows part of the data of interest, all towns from "Antwerpen" (part of Vlaanderen) with the latitude, longitude gps coordinates of the town centers:



The following selection criteria are applied:

* Select data where City = Antwerpen
* Remove duplicate towns with identical gps coordinates (eg. town of Antwerpen is divided in different districts with different postal codes). This will leave us with 77 towns and the idea is to form 3 groups of towns. Each group will be different with respect to the number of venues related to coffee/breakfast shops as well as shopping area’s, public transportation spots or other (touristic) attractions.

Additionally, a variety of Python packages will be used to analyze and display the data to gain insights and eventually provide a recommendation.

# Methodology

We will use the Foursquare data in combination with above location data to retrieve the number of similar places (venues) and the attractions in the direct neighborhood. Then we will use KMeans clustering to divide the towns of "Antwerpen" in 3 clusters. The idea is to open one new shop in each of the 3 segments.

Some exploration on the data has been done to identify all the unique categories of venues returned by the Foursquare explore API. Then it was decided to keep only the venues of interest, namely Coffee Shops and Breakfast Spots as our target categories and all categories involving shops, stores, public transportation and town attractions like theaters, zoo’s, etc.

Further exploration of the data exposed that 75% of the towns didn’t have more then 10 venues belonging to the categories we’re interested in. So it was decided to keep only the towns with at least 10 venues of these categories. This left us with 20 towns.

# Results

The KMeans partition based clustering algorithm divided the towns in 3 segments, respectively containing 4, 3 and 13 towns. We will call these partitions Cluster 1, Cluster 2 and Cluster 3 onwards.

All the towns in Cluster 1 have a “coffee shop” as the top venue in a radius of 1500m around the town center (identified by the latitude and longitude coordinates). This uniquely classifies this cluster. The other main venues are all shops/stores and some attractions, but no public transportation stations.

The towns in the second cluster (Cluster 2) also show some coffee shops (2 out of 3) but all rank 8th or less in the ranking of all relevant venues. So this cluster is clearly different from the first cluster and this is exactly what we are looking for.

The Cluster 3 towns are then again different from the Cluster 1 and 2 towns as in only 25% of the towns a coffee shop is in the top 10 of venues. But all (except 1) are ranked with lower importance (spread between the 2nd and 9th position).

in the top 10 venues of Cluster 2 and 3. Most of the towns in Cluster 2 have public transportation stations in the neighborhood of the center. None can be found in Cluster 1 and 3.

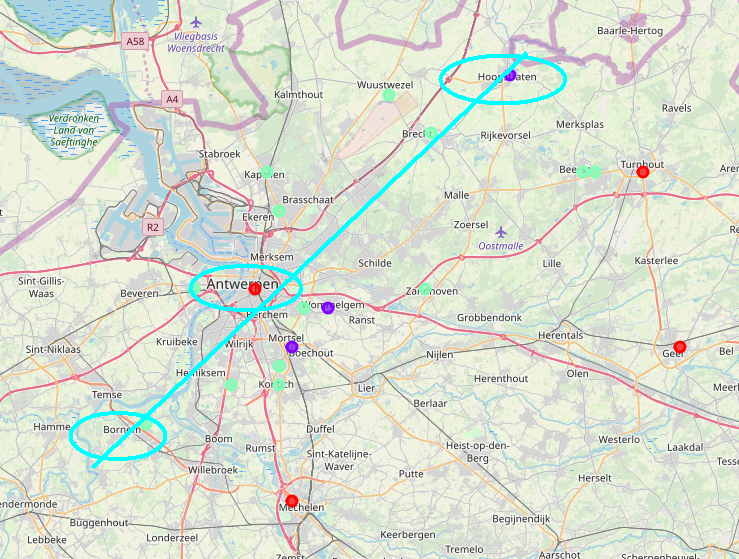
We can also note that only Cluster 3 has public transportation stations in their top 10 ranking of venues whereas all clusters show limited (touristic) attractions.

The geographic spread (shown in the chart) does not make us any wiser but gives a good view on where to open the 3 new shops as the idea is to start the shops in significant different locations. It does look like the towns of Cluster1 are the bigger ones and most of the towns of Cluster 3 are more centered around the Anwerp province capital, also called Antwerp.

# Discussion

It is obvious that our recommendation to the investor would be to open a new coffee/breakfast shop in one of the towns of each cluster identified.

Hoogstraten-Antwerpen-Bornem could be a good choice as these towns are geographically far away from each other. On the other hand it could be wise to choose for another town (like Mechelen or Geel) in Cluster 1 as Antwerpen probably has a higher risk in regards to competing with more similar shops.



Cluster 1: red circles

Cluster 2: purple circles

Cluster 3: green circles

Hoogstraten would be the most recommended spot in Cluster 2 as it is more isolated from the rest of the towns included in this analysis.

The decision where to start a shop in Cluster 3 can’t be positively identified based on this analysis and would require further market analysis. Maybe it should be recommended to start a new shop in the town which is geographically the furthest away from the towns chosen for Cluster 1 and 2.

# Conclusion

The KMeans classification technique gives the investor a good idea how and where to spread his new shops, based on the type of venues identified by the Foursquare api. However this type of analysis does not uniquely identify the best town to start a new shop but it gives a pretty good idea of how the province of Antwerp can be segmented based on the criteria put forward by the investor. This will prevent the investor of opening 2 shops in the same type of neighborhood, having the same type of venues nearby.