# Capstone project

### Problem & background

Recently I visited Gent, one of the biggest and most visited cities in Belgium. I was wondering how similar or dissimilar this city is from the city I live in Antwerp. From visiting Gent, I felt it was a bit like Antwerp, but I want to check if this feeling is right.

The analysis of this problem can be of interest to those who want to choose between visiting either Antwerp or Gent. The **stakeholders** will typically be tourist that are visiting Belgium or Europe and do not have time to visit both cities.

### Data

We can compare these cities by clustering their neighborhoods/areas by type of venues, by using the foursquare API, and comparing the resulting clusters. The API offers real-time access to Foursquare’s global database of rich venue data and user content. We can also use the foursquare API to compare to most common venues in both cities.

To obtain the neighborhoods I used online sources and my own knowledge of the cities to make an Excel file that contains different neighborhoods/areas. I limited the scope to the more inner parts of both cities (see later in the result section).

### Methodology

The main methodology that was being used was clustering and more specific: k-means clustering.

K-means clustering aims to partition several observations into k clusters in which each observation belongs to the cluster with the nearest mean. K-means clustering tries to minimize the intra-cluster variances or the so-called inertia. The different observations are in this case the different neighborhoods or areas. Each observation contains a lot of features: relative amount of each type of venue within a certain radius of the ‘center’ of the neighborhood/area.

Obtaining the optimal number of clusters is a fundamental issue in partitioning clustering. The determination of the optimal amount is somewhat subjective.

The elbow method has been used to check if we can determine an amount that is somewhat optimal.

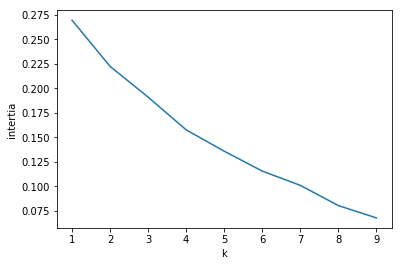
In the elbow method one should choose the number of clusters such that adding another cluster does not result in much better minimization of intra-cluster variation or inertia.

For both cities, Antwerp and Gent, the optimal number has been chosen this way.

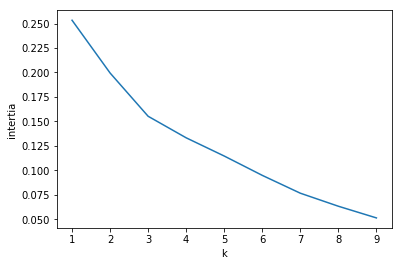
**Figure** 1 shows the elbow graph for Antwerp. It is not so easy to determine the optimal number of clusters here: 2 as well as 4 clusters seem to be viable candidates. This is the reason why both options were considered and further explored.

**Figure** 2 shows the elbow graph for Gent. Here it is much easier to determine the optimal number of clusters. After k = 3 the graph seems to decrease much more slowly.

**Figure** 1: Inertia in function of the number of clusters k for **Antwerp**. Optimal number of clusters seems to be 2 or 4.



**Figure** 2: Inertia in function of the number of clusters k for **Gent.** Optimal number of clusters seems to be 3.



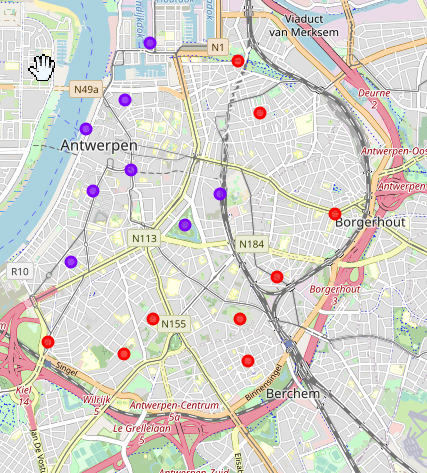
For further comparison, the 10 most common venues of Antwerp and Gent were obtained.

### Results

#### Antwerp

First, we look at the results of Antwerp. Here we clustered neighborhoods/areas using 2 as well as 4 clusters (see also methodology). **Figure** 3 shows the result of the clustering.

**Figure** 3: Results of clustering areas in Antwerp. K = 2.



It is not surprising that using 2 clusters, clustering the neighborhoods in Antwerp result in these two clusters. Since I am a resident of this city, I can use my "domain knowledge" here.

The second cluster (purple) is the most touristy part of the city there are the most diverse venues: zoo, coffee shop, cocktail bars, restaurants, shops, hotels, see .**Figure** 4

The first cluster is a more residential part of the city. The diversification of venues is much less than the other cluster and there are also not as much clothing stores. In this cluster there are an especially large number of bakers and bars, see **Figure** 5.

**Figure** 4:Most common venues grouped by neighborhood/area in **Antwerp** for the purple cluster.



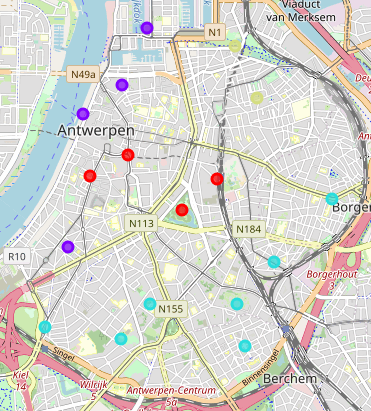
**Figure** 5: Most common venue grouped by neighborhood/area in **Antwerp** for the **red** cluster.



It is remarkably interesting to see that if we increase the number of clusters from 2 to 4 clusters, we just get a subdivision of the first two clusters.

The cluster of the more touristy part of the city gets split in clusters 1 (red, **Figure** 7) and 2 (purple, **Figure** 8). The red clusters seem to be the area that is most suited for shopping. For the less touristy clusters it is harder to determine the meaning of the subdivision. Although if you want to get a drink, you are far better off in the third cluster (blue, **Figure** 9).

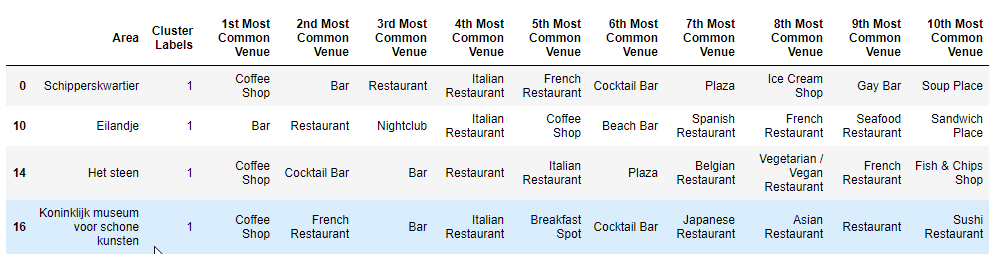
**Figure** 6: Results of clustering areas in **Antwerp**. K = 4.



**Figure** 7: most common venues grouped by neighborhood/area in Antwerp for the **red** cluster.



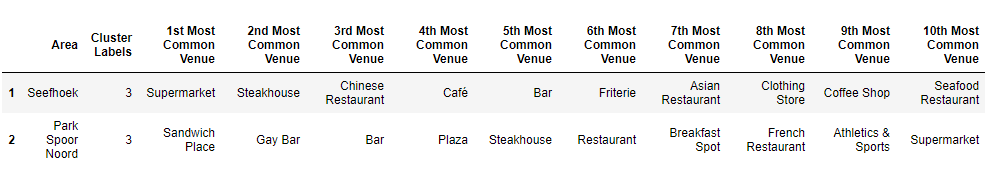
**Figure** 8: most common venues grouped by neighborhood/area in Antwerp for the **purple** cluster.



**Figure** 9: most common venues grouped by neighborhood/area in **Antwerp** for the **blue** cluster.



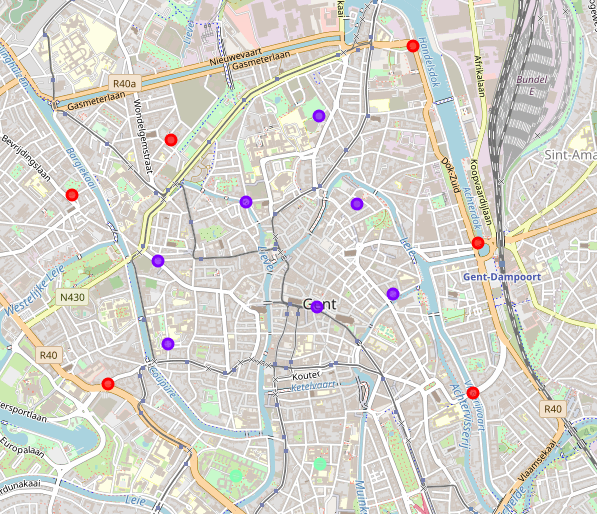
**Figure** 10: most common venues grouped by neighborhood/area in Antwerp for the **yellow** cluster.



#### Gent

If we now repeat the same process for **Gent** with the optimal number of clusters (k = 3) we find the clusters depicted in. **Figure** 11.

**Figure** 11: Results of clustering areas in **Gent**. K = 3.

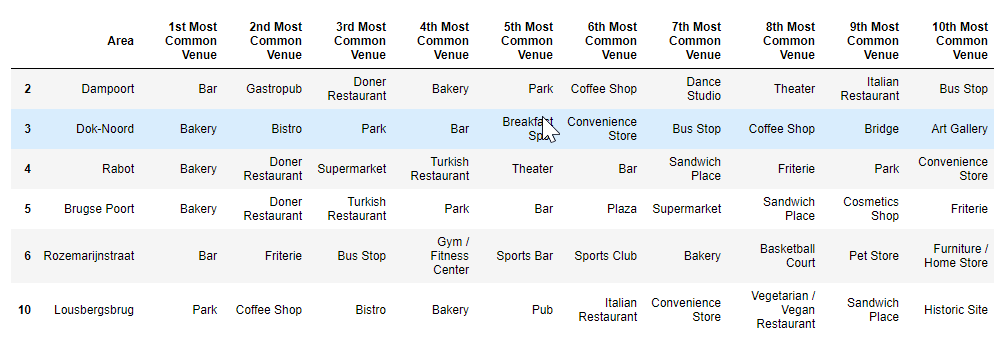


A similar pattern as with the clustering for Antwerp arises: a more touristy part can be found around the center of the city: the purple cluster (see **Figure** 12 ) while other parts, the red (see **Figure** 13) and the green cluster (see **Figure** 14), are more residential: a lot of bakers and bars and less shopping venues. The difference between the red and the green cluster can be found in a that the green cluster is dominated by pubs and sandwich places.

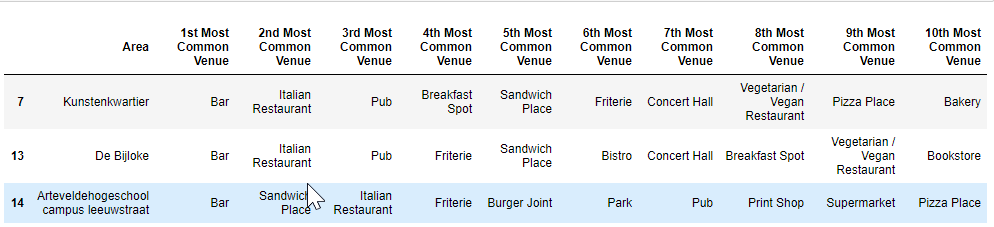
**Figure** 12: most common venues grouped by neighborhood/area in **Gent** for the **purple** cluster.



**Figure** 13: most common venues grouped by neighborhood/area in **Gent** for the **red** cluster.



**Figure** 14: most common venues grouped by neighborhood/area in **Gent** for the **green** cluster.



#### Further comparison

For a final comparison let us look at the most common venues in Gent and Antwerp.

Both have coffee shops as the most common venues. Gent seems to be more of a fashion city compared to Antwerp: clothing store is the second most common venue in Gent while in Antwerp it is not even in the top 10. Both cities seem to have a lot of diverse restaurants.

Antwerp seems to have relatively more bars than Gent while Gent has more Cocktail bars.

Figure : most common venues in **Gent**.

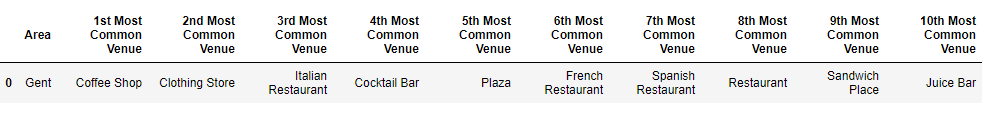
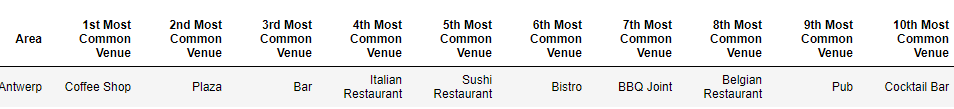


Figure : most common venues in **Antwerp.**



### Discussion

Antwerp and Gent indeed seem to have sort of the same pattern of type of neighborhoods/areas in terms of venues: more touristy areas in the center of the city vs more residential areas in the periphery. Based on the clustering I can recommend visiting the centers since these center areas are a lot more diverse.

The difference between the city in terms of venues are

* Relatively more clothing stores in Gent while relatively more plaza’s and bars in Antwerp. For people that like the more small-scale shopping experience I would recommend Gent, while people that like the more shopping center experience I would recommend Antwerp.
* More bars in Antwerp: If you would like to hang out in the pub or bar, I would recommend Antwerp.
* A lot of sushi restaurants in Antwerp.
* Relatively more lunch opportunities in Gent.
* Relatively many juice bars in Gent
* Relatively more Spanish restaurants in Gent.

But overall, both cities are indeed remarkably similar.

### Conclusion

We found a similar cluster pattern and quite a lot of similar venues in the top 10 for Antwerp and Gent. Some small differences can be found and may turn the decision of a potential reader towards a certain city. However just throwing a dice might also work if the differences are of no interest to the potential reader.

If we want to research further differences in experience it can be interesting to take other type of features in consideration, e.g. different types of surface areas, population, amount of tourists, …

To make a recommendation for one or another city solely based on venues might not be enough for a lot of individuals.