**Capstone Project - The Battle of Dutch Cities**

**Applied Data Science Capstone Project by IBM/Coursera – I. Quint**

**Table of contents**

* [Introduction: Business Problem](http://localhost:8888/notebooks/Documents/Data%20Science%20Course/Capstone%20project/Capstone%20Project%20-%20The%20Battle%20of%20Dutch%20Cities%20-%205.1.ipynb#introduction)
* [Data](http://localhost:8888/notebooks/Documents/Data%20Science%20Course/Capstone%20project/Capstone%20Project%20-%20The%20Battle%20of%20Dutch%20Cities%20-%205.1.ipynb#data)
* [Methodology](http://localhost:8888/notebooks/Documents/Data%20Science%20Course/Capstone%20project/Capstone%20Project%20-%20The%20Battle%20of%20Dutch%20Cities%20-%205.1.ipynb#methodology)
* [Analysis](http://localhost:8888/notebooks/Documents/Data%20Science%20Course/Capstone%20project/Capstone%20Project%20-%20The%20Battle%20of%20Dutch%20Cities%20-%205.1.ipynb#analysis)
* [Conclusion](http://localhost:8888/notebooks/Documents/Data%20Science%20Course/Capstone%20project/Capstone%20Project%20-%20The%20Battle%20of%20Dutch%20Cities%20-%205.1.ipynb#conclusion)

**Introduction: Business Problem**

The purpose of this research is to compare all **provincial capitals** of the **Netherlands** on the basis of **facilities and venues**. Note how a few big Dutch cities aren't capitals within it's provice, so Rotterdam and Amsterdam will be included within the data sample as well. The analysis will be specifically based on the **relative frequencies** of venues per city inhabitant.

Dutch inhabitants often have conceptions about how for example Amsterdam has a large number of trendy coffee bars. How does this hold for our **Northern cities** if you correct for the number of inhabitants?

Using this analysis, several insights can be obtained. We will try to cluster cities based on the relative venue frequencies. As such, one can find comparable cities within a given radius. This can be used for various purposes, including population research, marketing and business propositions.

**Data**

Based on definition of our problem, factors that will influence our classification are:

* List of capital cities per province
* Number of inhabitants in the city
* Venues in (capital) city
* Exact city location (GPS coordinates)

Following data sources will be needed to extract/generate the required information:

* List of capital cities per province can be obtained from the **Provinces of the Netherlands Wikipedia page**
* Number of city inhabitants will be extracted from each cities' **Wikipedia page**
* Venues in (capital) city, their type and location will be obtained using **Foursquare API**
* Exact city location (GPS coordinates) will be obtained using **Geopy package**

**Capital Cities**

To retrieve the Dutch province capitals and its inhabitants, we use a Wikipedia page and extract the data. Each cities’ Wikipedia page is searched for the number of inhabitants, and this is added to the dataframe. Table 1 presents the final result.

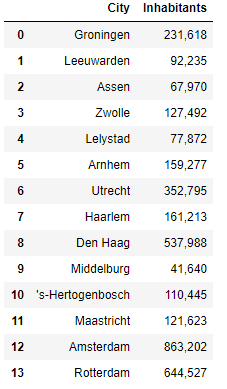


Table 1. Cities in sample

## Methodology

The analysis will be done by retrieving venues for each of the cities and getting all the venue categories. The number of venue per category per city will be divided by the cities inhabitants divided by 1,000. Thus, our result will show the relevant frequency of each venue per 1,000 inhabitants. This will allow for easier comparison.

Next, using Scikit Learn, we will use the Kmeans clustering algorithm to determine the optimal number of clusters and run the analysis. Finally, these clusters will be marked on a map of the Netherlands, showing how cities compare.

## Analysis

To start the analysis, first we have to import several libraries that we will work with. We will need Scikit Learn for the K-means clustering and determining the optimal K. Geopy will be used to find coordinates of our cities. Folium for creating the map and matplotlib for adding markers to the map.

After getting each cities venues, we sort them based on frequencies. The result is shown in Table 2.

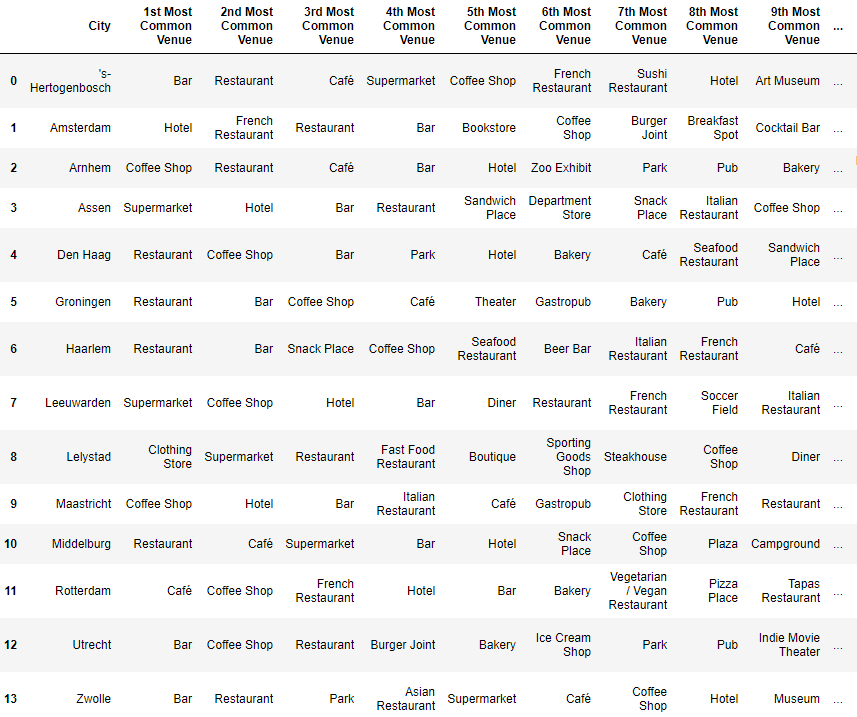


Table 2. Cities and their venues sorted on frequency

Now, the data has been prepped and we are ready to start analysing clusters within the data to determine cities that are alike. First, the optimal number of clusters is determined using the so-called elbow method. The result is shown in Figure 1.

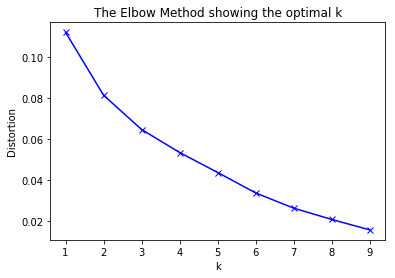


Figure 1. The elbow method chart

Figure 1 shows a sharp decrease in the cost function when k increases from 1 to 2. In the next steps, the cost function decreases only slightly between every step. Most notably, from 3 to 4 the decrease in the loss function is almost constant as from 4 to 5. This indicates us that with k=3, we have a good balance between number of clusters and the information. Hence, the K-means algorithm will be run again with our chosen number of clusters (k=3).

### Map

Next, we will plot all cities on the map and indicate the different clusters using coloured markers. See the result in Figure 2.

Figure 2. Map with cluster markers on each of the cities

## Conclusion

This analysis compared the capital cities of the provinces in the Netherland (and Rotterdam and Amsterdam included as well). Using K-means clustering, the optimal number of clusters was found to find similarities within the number of venues per 1,000 inhabitants within the Netherlands. As can be seen from the map, there is one big cluster with similar cities. This cluster consists of the following cities:

0 Groningen

1 Leeuwarden

2 Assen

3 Zwolle

4 Lelystad

6 Utrecht

7 Haarlem

9 Middelburg

10 's-Hertogenbosch

12 Amsterdam

13 Rotterdam

Name: City, dtype: object

Next, there is a much smaller cluster with only the following cities:

5 Arnhem

8 Den Haag

Name: City, dtype: object

Finally, a cluster of just one city remains:

11 Maastricht

Name: City, dtype: object

Interestingly, this last cluster is very different from the other clusters. As can be observed by the common venues, hotels and coffeeshop are very frequent for a city with a relatively low number of inhabitants. This may be explained by tourism.

City Maastricht

Inhabitants 121,623

Latitude 50.858

Longitude 5.69699

Cluster Labels 1

1st Most Common Venue Coffee Shop

2nd Most Common Venue Hotel

3rd Most Common Venue Bar

4th Most Common Venue Italian Restaurant

5th Most Common Venue Café

6th Most Common Venue Gastropub

7th Most Common Venue Clothing Store

8th Most Common Venue French Restaurant

9th Most Common Venue Restaurant

10th Most Common Venue Pub

11th Most Common Venue Plaza

12th Most Common Venue Ice Cream Shop

13th Most Common Venue Park

14th Most Common Venue Bistro

15th Most Common Venue Sushi Restaurant

16th Most Common Venue Tapas Restaurant

17th Most Common Venue Thai Restaurant

18th Most Common Venue Bakery

19th Most Common Venue Cocktail Bar

20th Most Common Venue Chocolate Shop

Name: 11, dtype: object