# Report on Dutch cities

By Stefan Arends as Part of the Capstone project ion “IBM Data Science Professional Certificate

# Introduction

I am planning to move to the Netherland, a small country in Europe. Because it is a small county, all cities are acceptable because of traveling distances. To have more insights of which city to select for moving to, I want to unlock the data available in Foursquare on analysing the different cities.

Based on the information I want to cluster the different cities with machine learning, so I have insight on similar cities within the Netherlands.

**Problem statement:**

Give insights in the characteristics of the 50 biggest cities in the Netherlands, based on venue information in Foursquare.

This can be input for selecting a (cluster of) cities which are candidate to move to.

# Data

For this analysis the following data sources will be used:

1. **List of Dutch cities,** downloaded from the CBS, the Dutch Statistics Agency. Link: <https://www.cbs.nl/nl-nl/maatwerk/2019/31/kerncijfers-wijken-en-buurten-2019>
2. **Geo coordinates** of the major Dutch cities, downloaded from <https://simplemaps.com/data/nl-cities>
3. **Top venues** in each Neighbourhoods gathered from Foursquare by the API.

These data sources will be combined, to cluster the different Dutch cities.

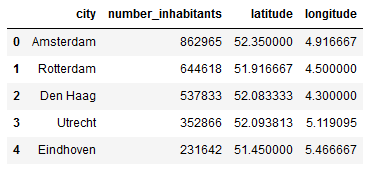
# Methodology

For the analysis the K-means clustering algorithm is used. In order to use this, several steps need to be taken:

* A list of 50 biggest cities need to be found. For this project a list of the CBS (Dutch statistics Agency) is available in excel.
* Because this does not contain geographical information about the location of the cities, this list must be combined with geographical information. This will be done by merging with another data source, containing latitude and longitude for around 400 Dutch cities.
* After combining the result is a list with the top 50 dutch cities, including latitude an longitude.
* With this geo-location, the Foursquare-api will be used to gather popular venues for each city.
* For each city the top 10 category of venues will be calculated, including the mean use of it. This top 10 catagories can be seen as a characterization of the city.
* This characterization can be input for the clustering the of cities, to asign similar cities to the same cluster.
* For this clustering the K-means machine learning is used.
* Based on the clustering the clusters an cities can be analysed and characterized, as input for making the choice.

# Results

First result is merging of the cities dataset with the geo dataset. The first 5 cities are presented in the table below



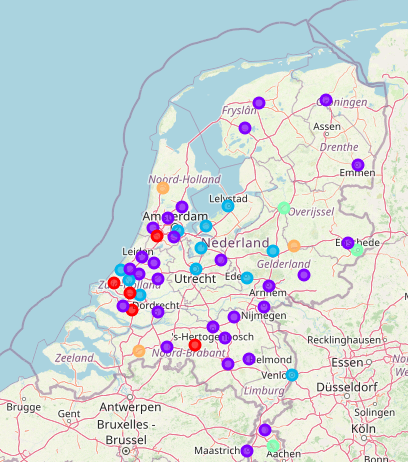
These cities can also be plotted on a map (without clustering yet)



Based on geo-location, for each city the top 10 venues cab be calculated, with result (only 5 first cities are shown)



By using K-means, this can be input for clustering the cities, with result below. Each color is representing another cluser.



For insight in the characteristics of (the cities in) the different clusters are shown by diving in the different clusters. This results are available in excel and also printed in the appendix



# Discussion

If you dive into the different clusters, there is not much difference shown between the different clusters. For all clusters the category “Office” is the number one for most cities because of the amount of the Foursquare venues in this category.

However small differences between the clusters can be identified.

A personal interpretation of the clusters can be:

* Cluster 0: Working environment with offices and automotive
* Cluster 1: Working environment with medical services in place
* Cluster 2: Working environment most time situated in the “Randstad” the most densely populated region of the Netherland
* Cluster 3: Studying environment
* Cluster 4: Living environment with furniture and home stores nearby.

# Conclusion

For selecting a city to live in, this analysis gives limited information, because the differentiation between the different cities and clusters in small. Maybe only cluster 3 and 4 differs from the others, based on there profile. So if Education (cluster 3) or Furniture/home stores (cluster 4) are important items, these cities can be considered.

To get more information on the cities, further research is recommended. See next section for recommendations on that.

# Recommendations

For gathering more information on the different cities, extra analysis is recommended. Possible options for that are:

* Filter out categories in Foursquare that are not relevant for making your choice (like “Office”)
* Use information in data source 1. A lot of information on the cities is kept in this dataset (like typical household, mean income, origins of inhabitants, age)
* Consider to unlock not only on city level, but also on Borough or Neighbourhoods.
* Use Google Places in stead of Foursquare, because of the low usage of Foursquare.

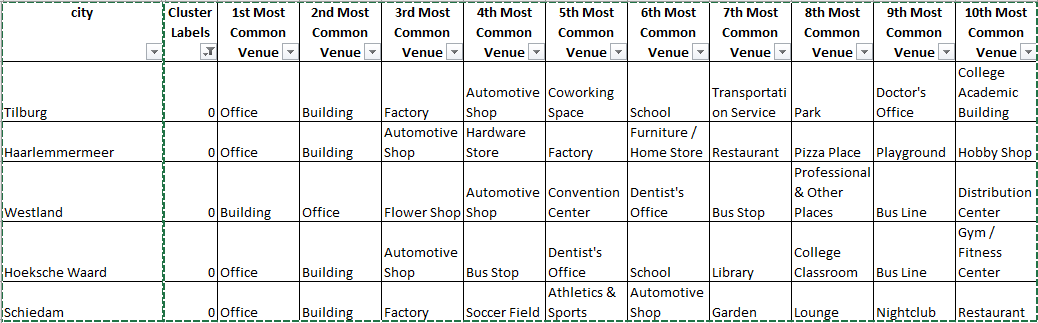
# Links

Gitbub location with all resources used: <https://github.com/StefanArends/CapStone>

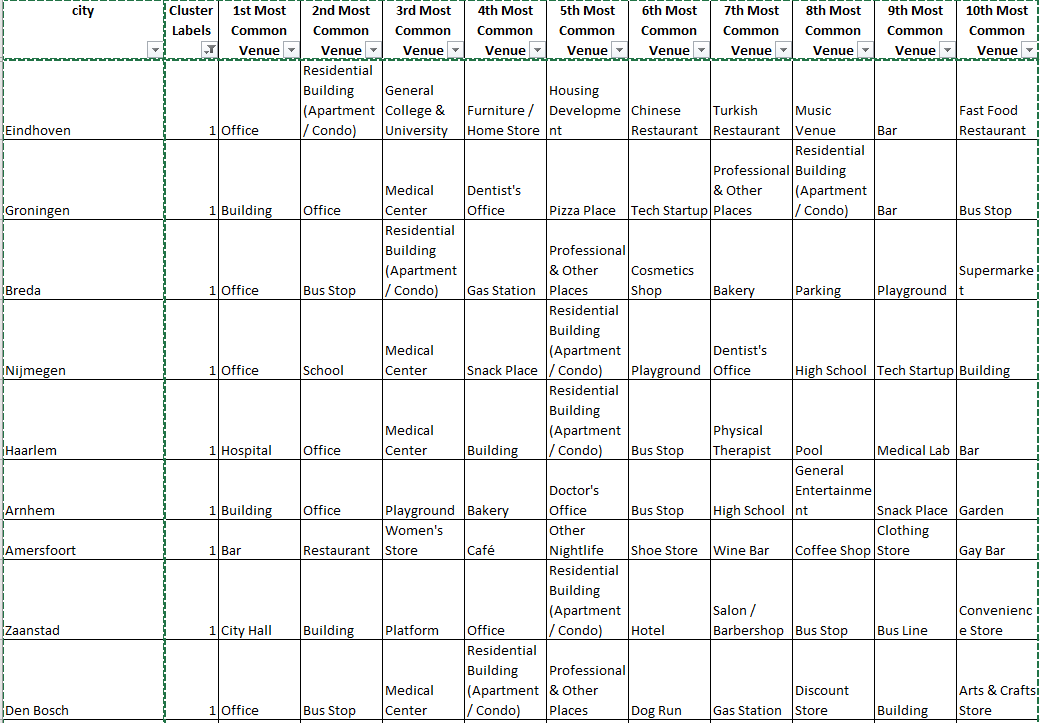
* Python Code in Notebook: <https://github.com/StefanArends/CapStone/blob/master/Clustering%20Dutch%20Cities.ipynb>
* Output exported to excel: <https://github.com/StefanArends/CapStone/blob/master/output.xlsx'>

# Appendix 1

## Cluster 0



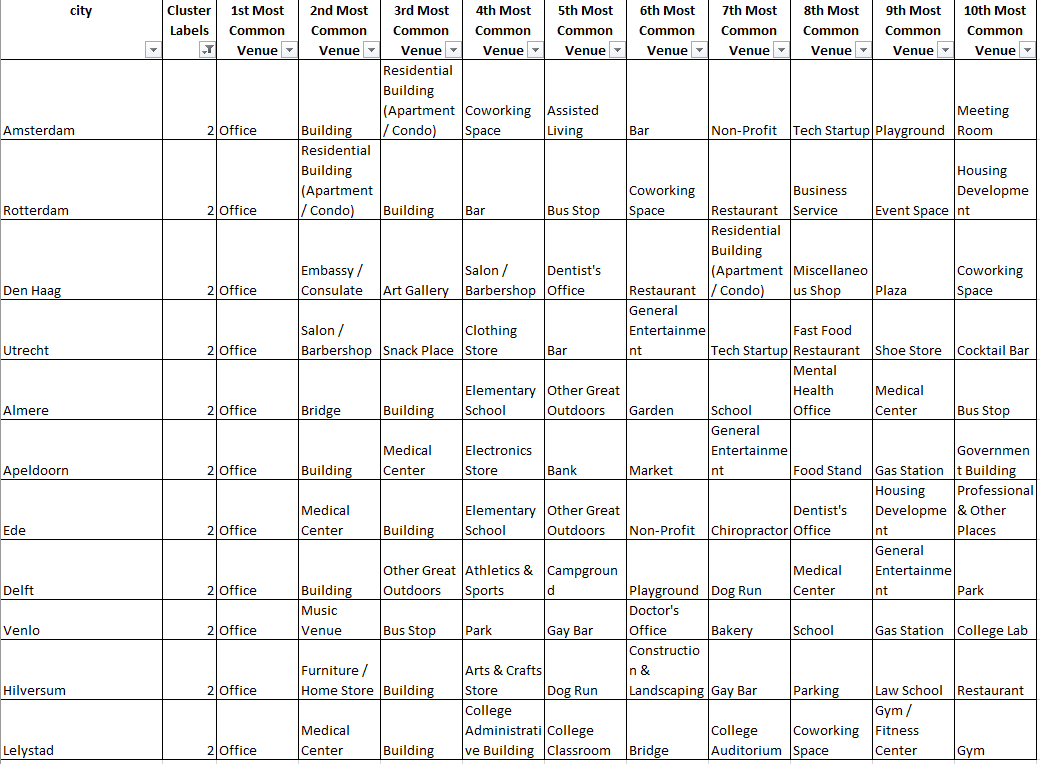
## Cluster 1



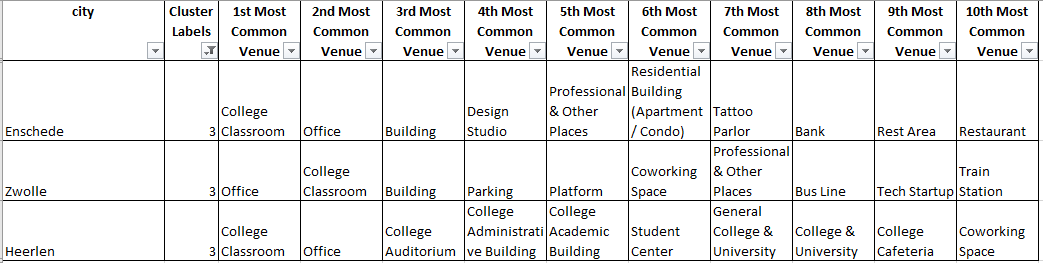
****

****

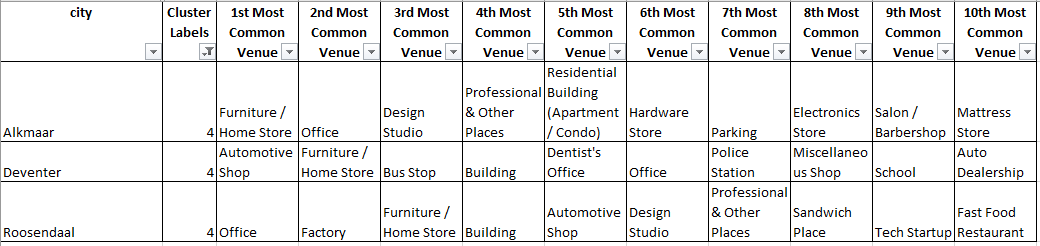
## Cluster 2

****

## Cluster 3

****

## Cluster 4

****