Classification and analysis of Dutch train stations

Arjan Lemmers

September 19th, 2019

'Capstone Project - The Battle of Neighborhoods' assignment, part of the Applied Data Science Capstone course on Coursera.

Table of content

This presentation is part of the 'Capstone Project - The Battle of Neighborhoods' assignment, part of the Applied Data Science Capstone course on Coursera.

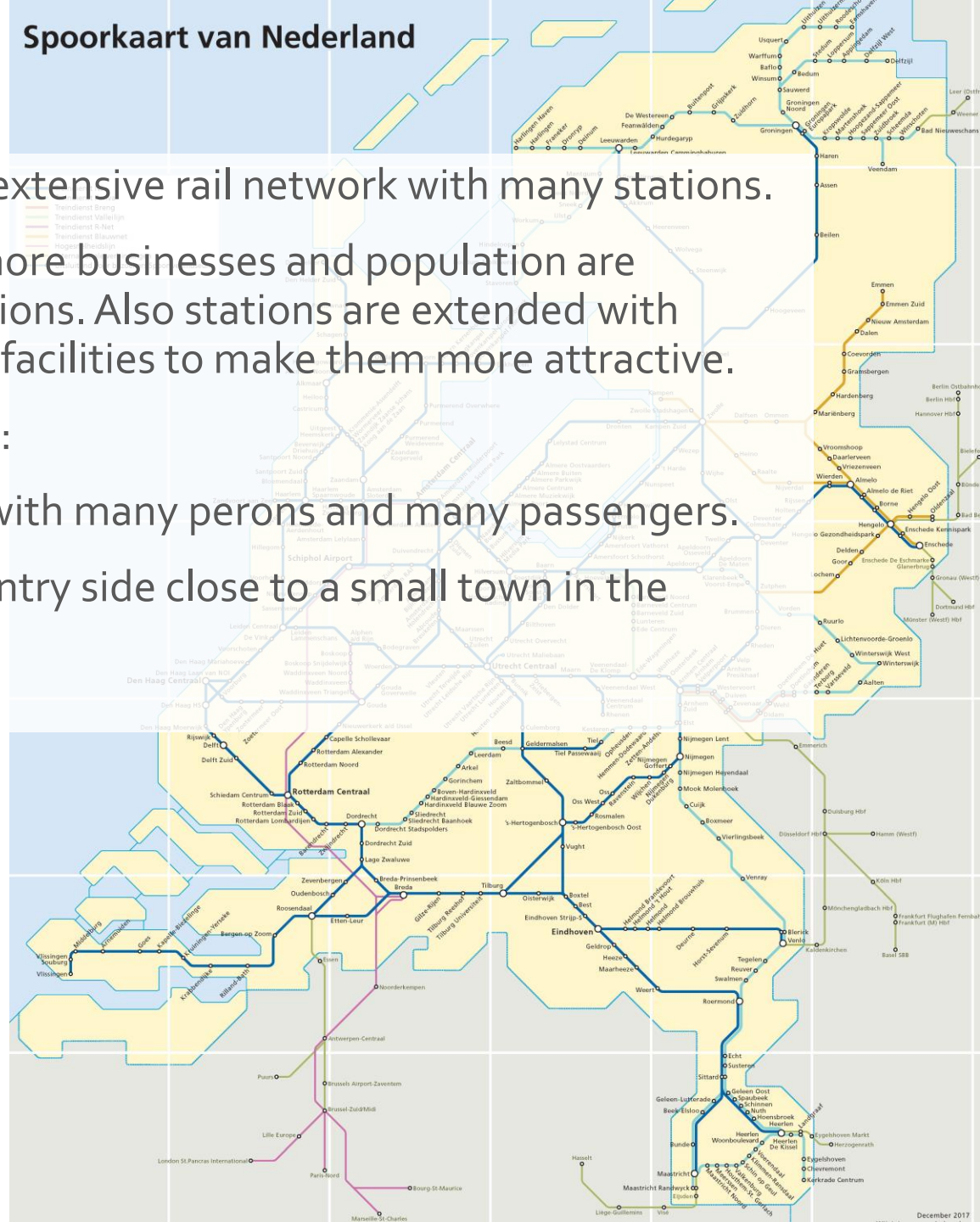
Introduction

The Netherlands have an extensive rail network with many stations.

The last years more and more businesses and population are concentrating around stations. Also stations are extended with shops, catering and office facilities to make them more attractive.

Stations are very different:

- Major hubs in big cities with many perons and many passengers.
- Local stations in the country side close to a small town in the middle of the fields.



Business Problem

The question is what are good candidate stations to invest in facilities, like shops or catering facilities.

The answer to this question helps an business investor to make the right investment decisions.

The goal is analyse and cluster the train stations in the Netherlands based on the characteristic of the stations and their neighbourhood.

Data Sources (1/2)

The required data is:

1. the Dutch train stations and their coordinates
2. the facilities at and around these train stations
3. the population per postal code (=zip code)
4. the coordinates of the Netherlands
5. the postcode for each station.

For point 1)

The NS is the major Dutch train operator. They provide the NS API (https://www.ns.nl/en/travel-information/ns-api) .

This API provides an Station section with endpoints to:

- get a list and coordinates of the train stations
- get details per station.

Data Sources (2/2)

Regarding point 3)

The goverment's CBS (Central Bureau Statistics) provide various data endpoints, including 'Population and households per postal code: https://beta.opendata.cbs.nl/DataPortal/detail/CBS/82245NED .

Via this table we retrieve the population per postal code.

Regarding point 4) and 5)

the the Nominatim geocoder for OpenStreetMap data from python library geopy is used.

Methodology

Based on the business problem, I like to compare the stations by clustering.

I like to cluster and categorize the stations on:

- number of venues around the station
- number of shops at the station
- the population close to the station.

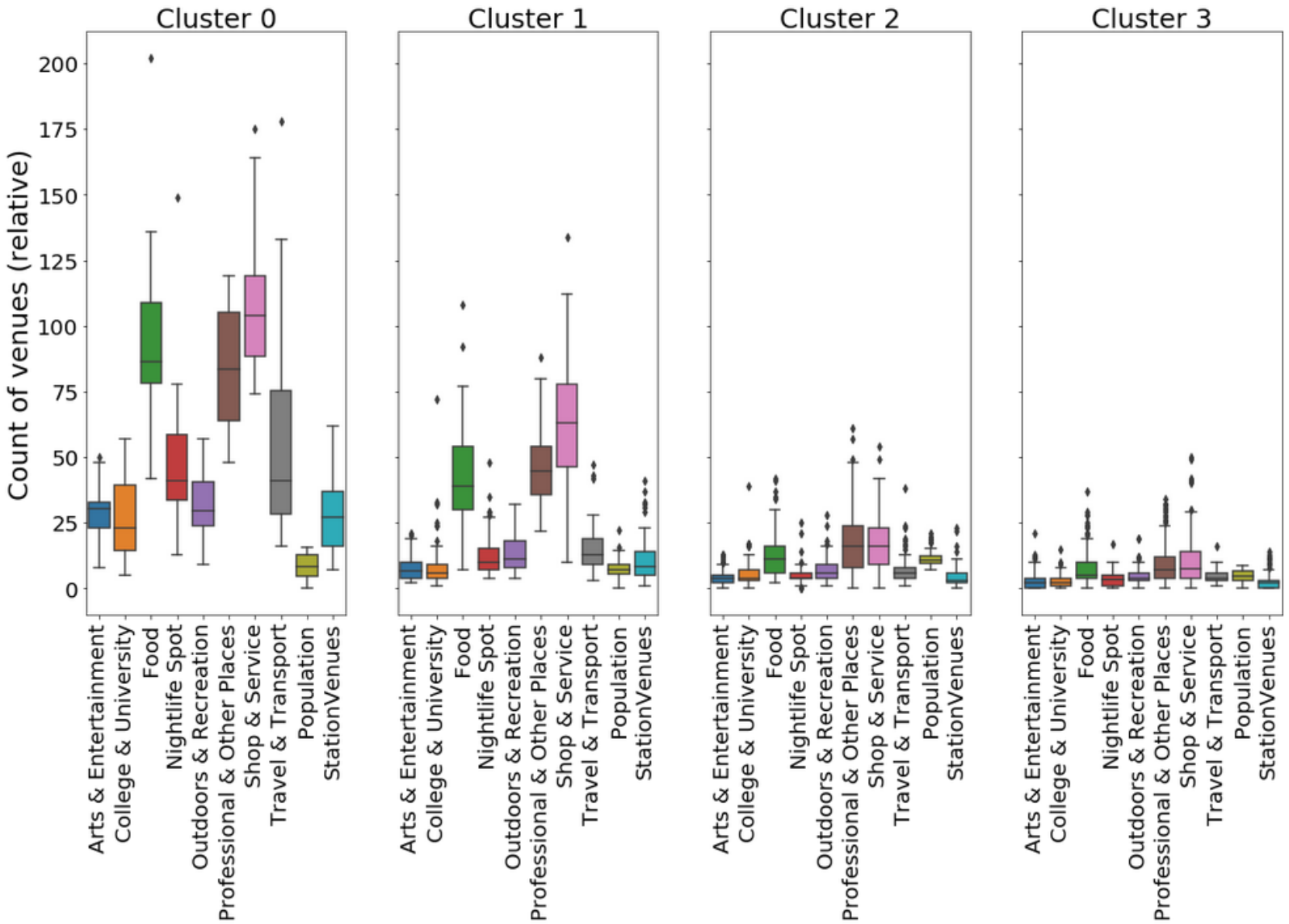
This categorization gives insights the population and activities around stations and select stations for a in-depth analysis for investment opportunities.

Remark:

> For deeper analysis of business opportunities I searched for data sources providing the number of passengers per station (per day). However this data was not available. As alternative I explore the population per station's postal code.

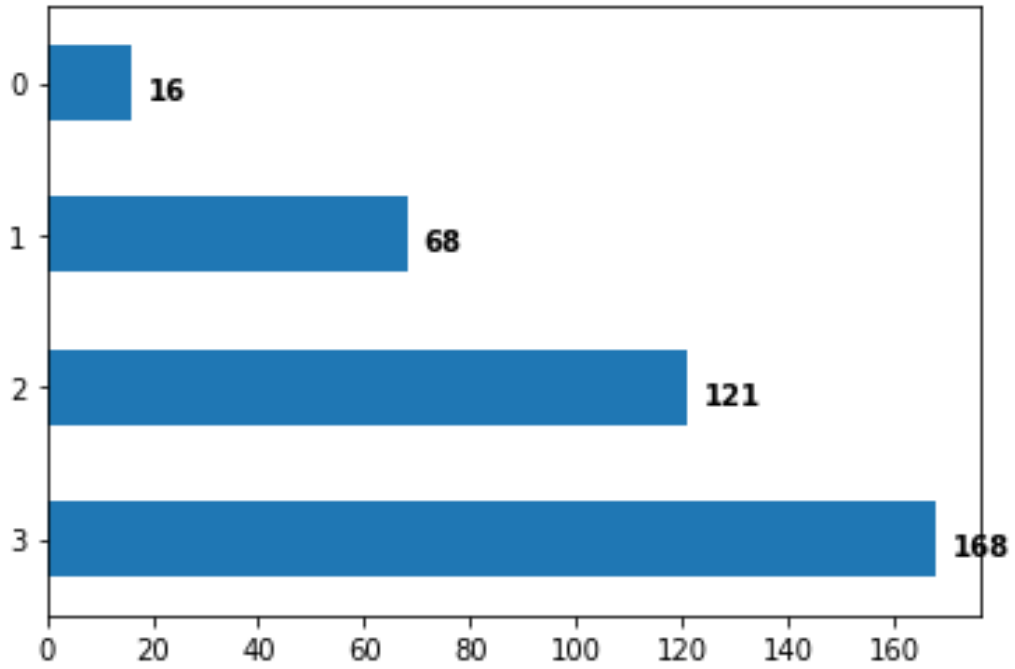
Data clustering (1/4)

K-Means clustering into 4 clusters



Data clustering (2/4)

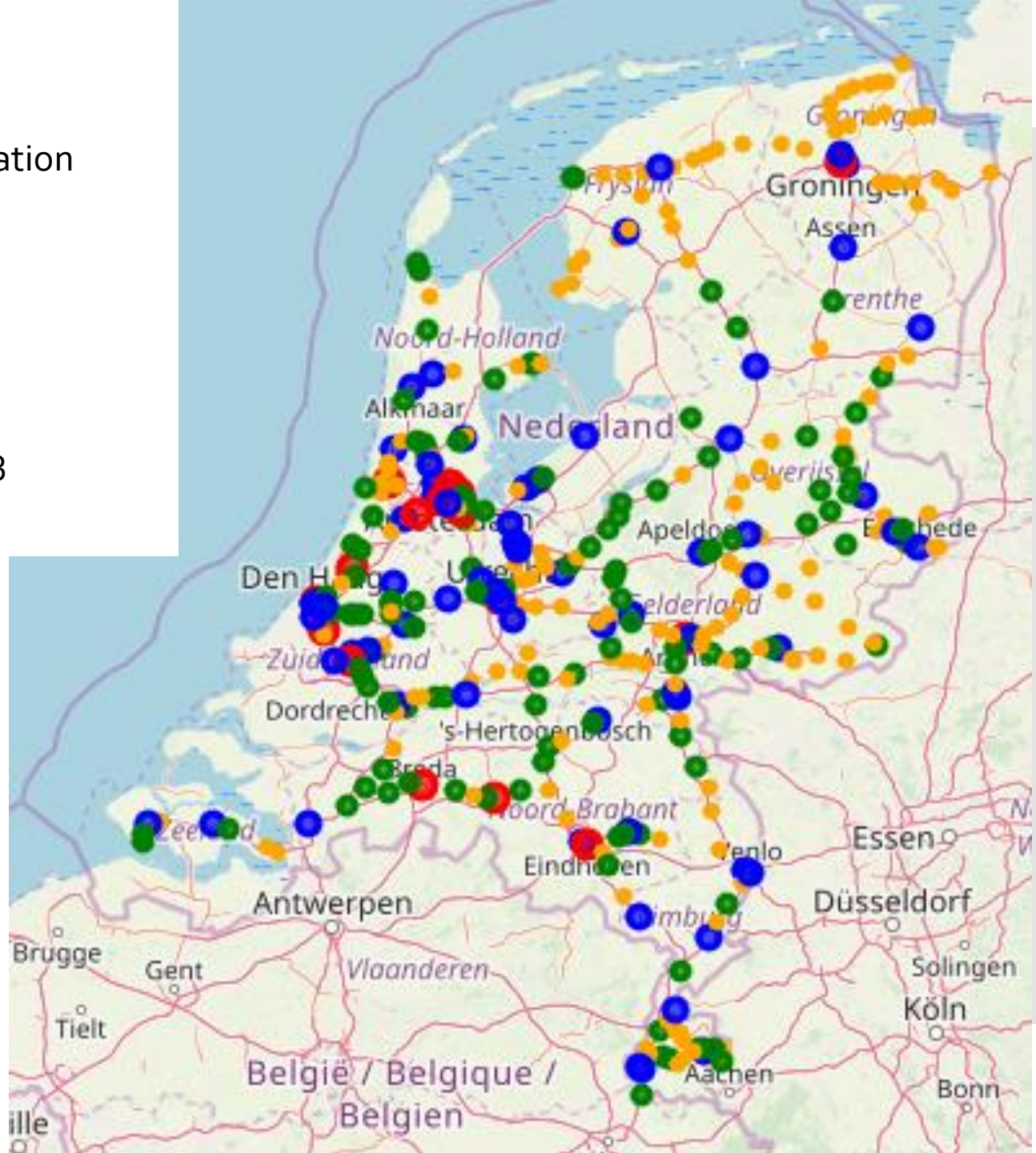
The number of stations per cluster:



Data clustering (3/4)

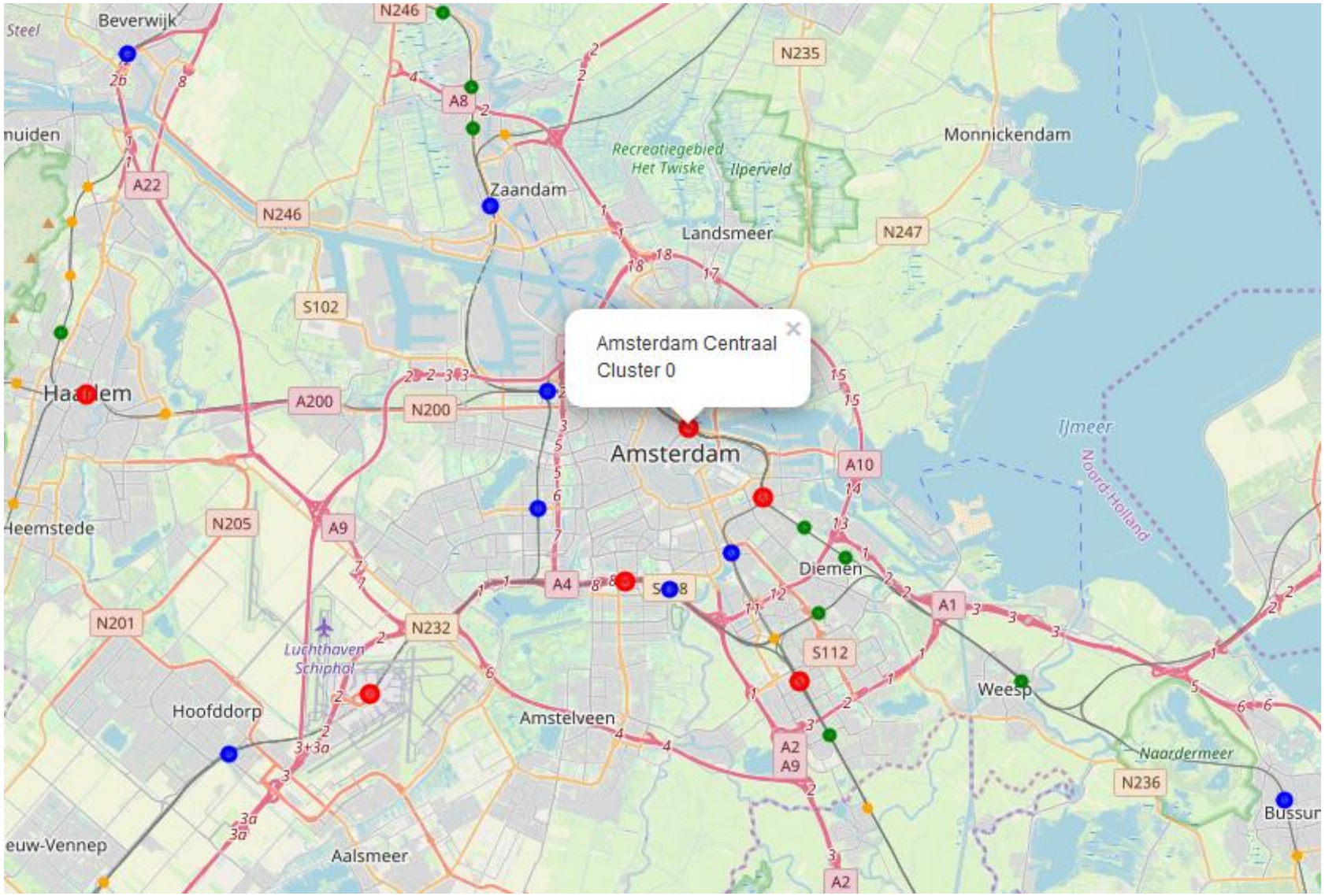
Geographical location stations:

Red – cluster 0
Blue – cluster 1
Green – cluster 2
Yellow – cluster 3



Data clustering (4/4)

Geographical location stations – Detail map around Amsterdam



Discussion

The current approach results into a clear clustering of number of venues around the station.
The smaller stations shows that the majority of venues are of category 'food', 'Professional & Other Places' and 'Shop & Service'.
The big stations shows a more spread pattern.

I tried to find any data regarding the numbers of passengers per station. I was expecting that the ration between venues on the station and the number of passengers would provide some indication stations with investment opportunities.

As alternative I tried the population around the stations. However to my surprise the population (in the stations postal code) provides no additional information. For all categories the population number and spread are very similar.

I noticed that some major stations are filtered out the data. To make the data more complete it must be investigated why the data is filtered out: for example missing coordinates or duplicate station codes.

Conclusion

Foursquare data is provides good insights to cluster the stations on its importance and size based the number of venues around the station.

However additional information are required to get more business insights in the stations:

- number of (daily) passengers per day per station. Passengers transferring or passengers entering of leaving the train system.
- rental prices of commercial spaces up or around a station..