Classification and analysis of Dutch train stations

Author: Arjan Lemmers
Date: September 19th, 2019

Table of content

Introduction

This report covers an assignment for the Coursera course 'Capstone Project - The Battle of Neighborhoods' assignment, part of the Applied Data Science Capstone course.

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.

Figure 1 Railmap Netherlands

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

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.

Regarding 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.

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 I retrieve the population per postal code.

Regarding point 4) and 5)

I use the Nominatim geocoder for OpenStreetMap data from python library `geopy` for coordinate related operations.

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 an 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

The following data was retrieved per station:

- The venues of the following categories are retrieved within 1000 meters of the station:
    - Arts & Entertainment
    - College & University
    - Food
    - Nightlife Spot
    - Outdoors & Recreation
    - Professional & Other Places
    - Shop & Service
    - Travel & Transport
- The 'StationVenues' value which are the count 'Food' and 'Shop & Service' venues with 200 meters of the station.
- The 'Population' venue, which is the population – divided by 1000 – in the postal code of the station.

The boxplot how a stations mainly differ on 'Food', 'Profesional & other Paces' and 'Shop & Service' venues.

The 'StationVenues' distribution is smaller, meaning the number of shops in the station a 'restricted', with a few (bigger) stations as exception.

The distribution of the 'Population' is very limited, meaning the population around the station does not provide much additional information.

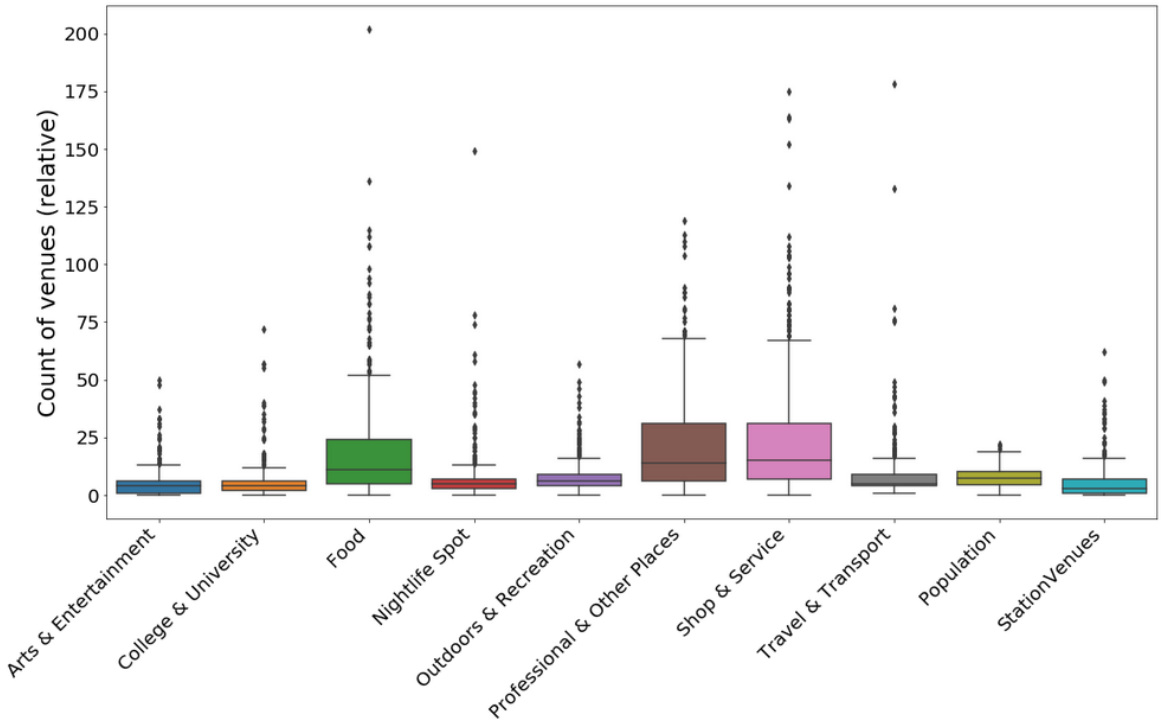

Figure 2 Boxplot non-normalized data

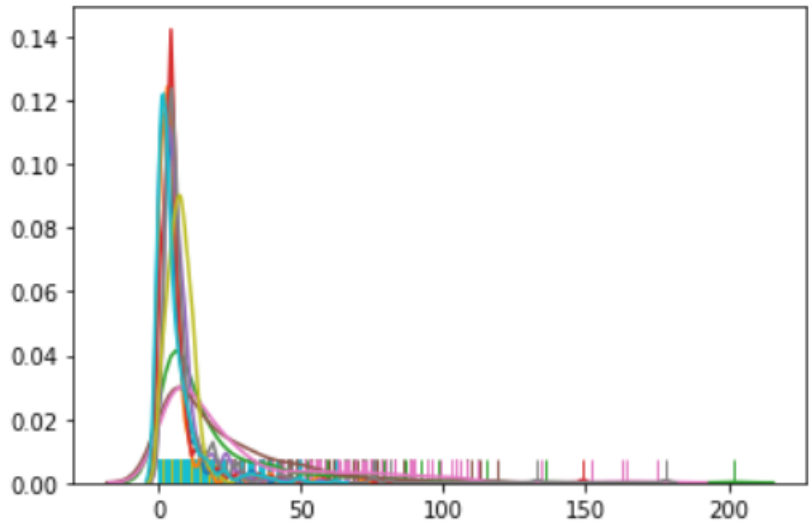

Figure 3 Distribution plot of the station properties

The data is normalized using MinMaxScaler: transforms features by scaling each feature to a given range. This estimator scales and translates each feature individually such that it is in the given range on the training set, e.g. between zero and one.

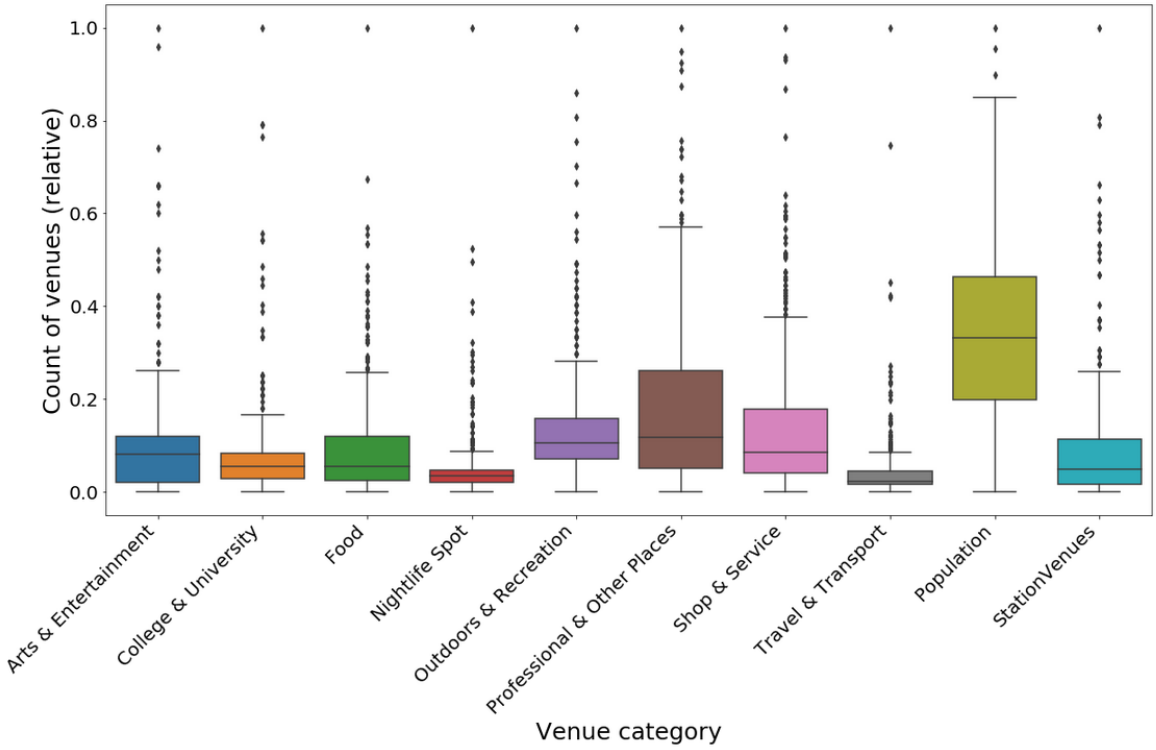

Figure 4 Boxplot of normalized data

To get more insights in the stations I use K-Means clustering. The stations are grouped in 4 clusters.

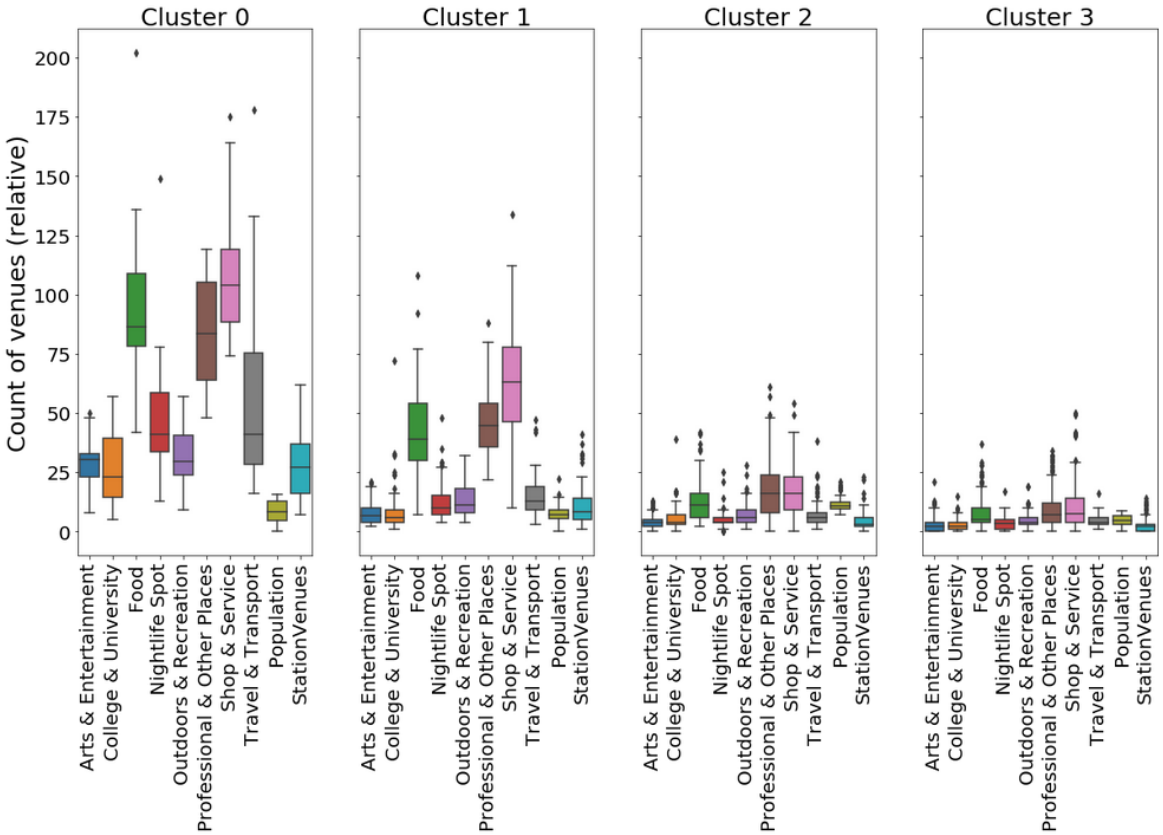

Figure 5 Boxplot of normalized data:

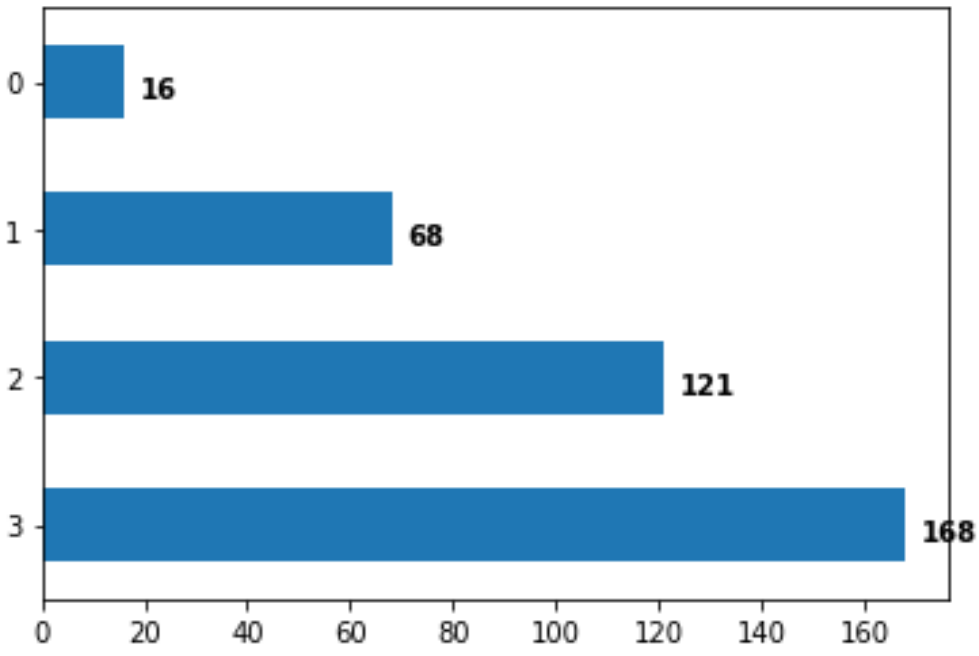

Figure 6 The number of stations per cluster

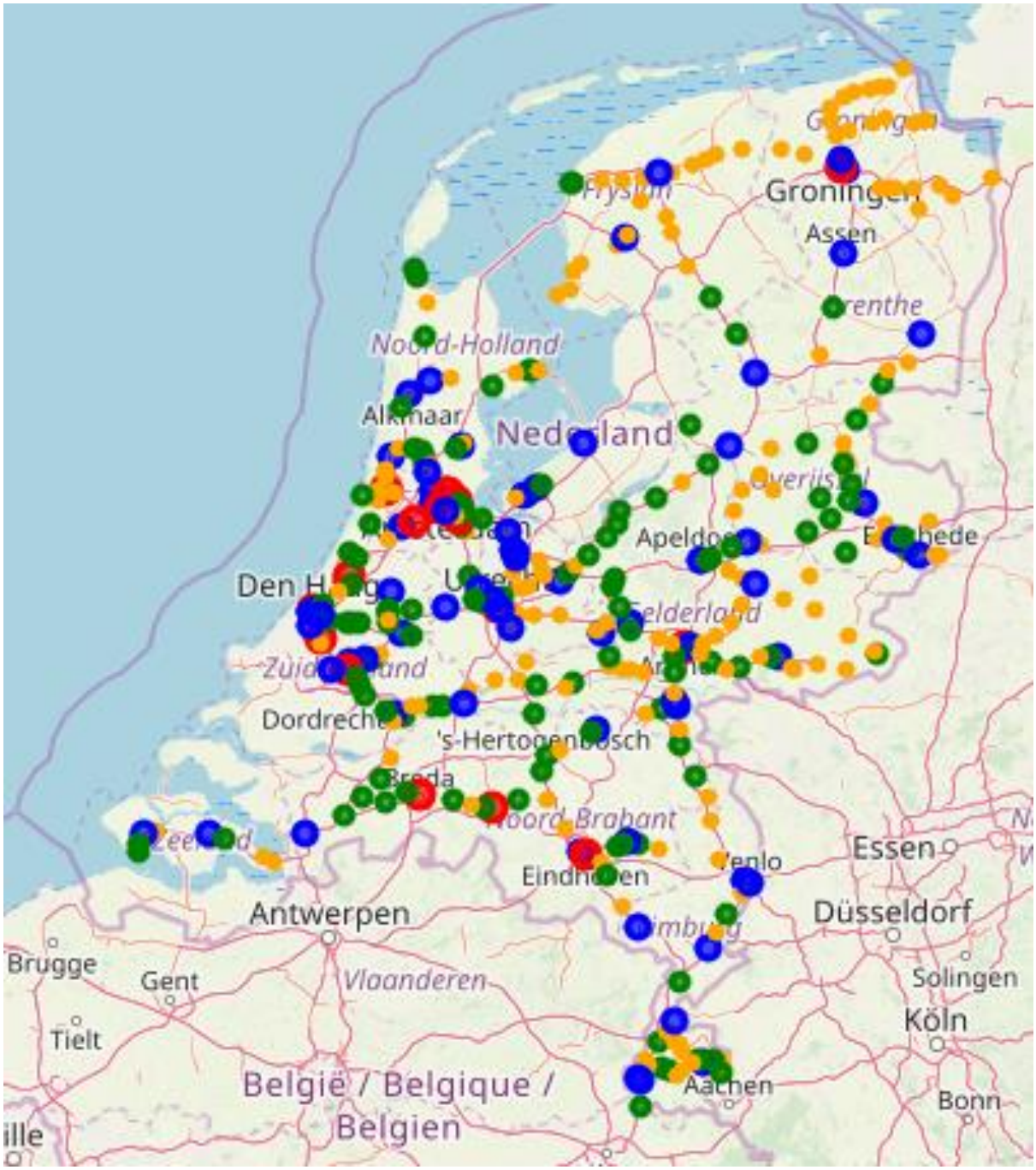

Figure 7 Geographical location

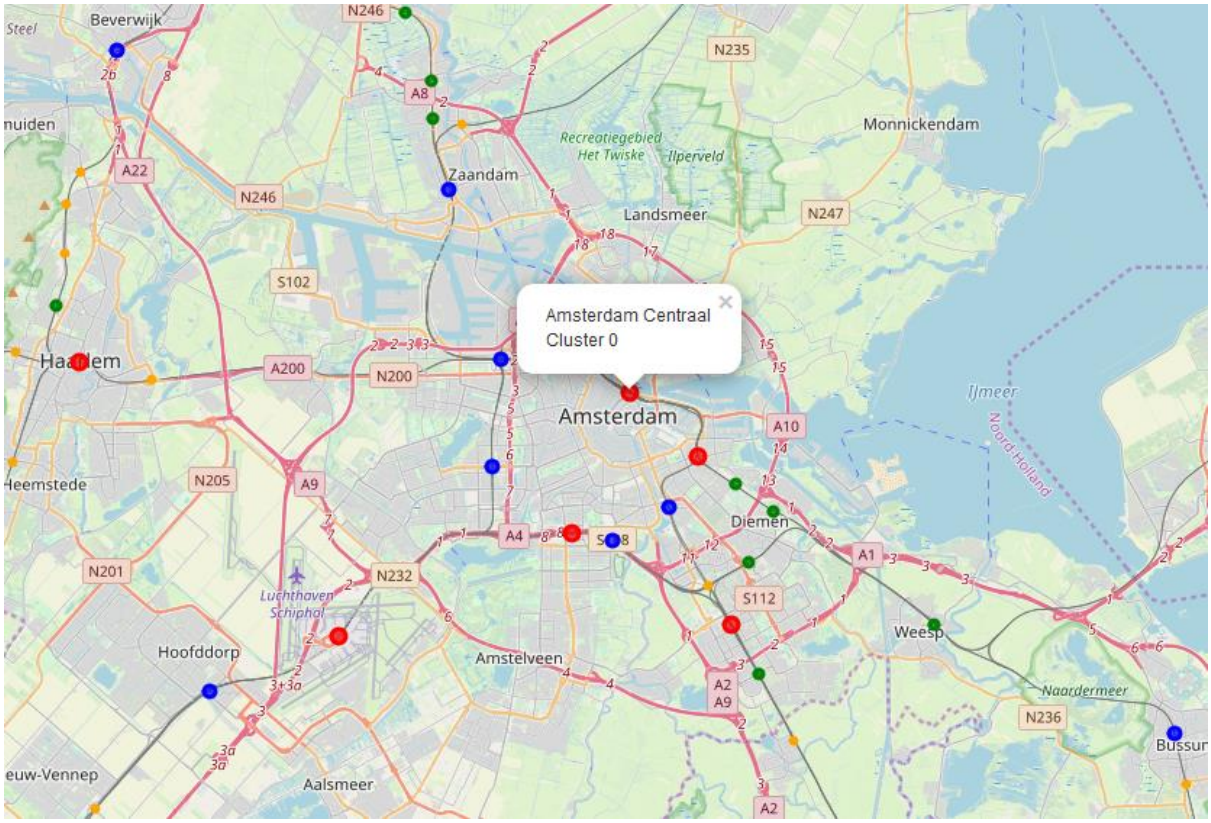

Figure 8 Geographical location stations – detail map around Amsterdam

The cluster groups the station on number of and type of venues in and around the stations:

- cluster 1 (Red) are the biggest and main stations: large transitions hubs in big cities with a large amount of facilities.
- cluster 2 (Blue) are the big stations: important stations in bigger cities with a big amount of facilities.
- cluster 3 (Green) are smaller stations but with some facilities.
- cluster 4 (Orange)are small stations in small villages or cities with limited or no facilities.

The choice of the number of clusters of 4 results in 4 clear clusters. The clustering mainly results on grouping based on number of venues. Change the number of clusters did not result into different insights than the stations size.

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 and does not provide any additional insights.

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.