

# Finding the optimum spot for the Halewijn Award Ceremony in Haarlem, the Netherlands.

Applied Data Science Capstone Project by IBM/Coursera

CRW Korver, data scientist *in spe*

email author: [crwkorver@hotmail.com](mailto:crwkorver@hotmail.com)

December 11, 2019



## 1. INTRODUCTION

The [Halewijn Prijs](#) is a Dutch national award, intended for literary talent that, based on quality and irresistibility of his or her published work, deserves extra interest. The prize consists of a cash prize and a bronze small statue made by Dick van Wijk.

Recently, it was decided that the yearly Award Ceremony will take place in the city of [Haarlem](#), the Netherlands. Haarlem is capital of the province of North Holland and has 161,265 inhabitants.

To attain maximal public attendance, the ideal neighborhood would be in the vicinity of the railway station in order to facilitate traveling by public means in a car-crowded city like Haarlem. Furthermore, to maximize potential book sales, a location should be found that is in the near vicinity of book shops around. Finally, ample restaurants should be present here as well, to accommodate the audience (and the winner and jury members, who will dine together afterwards).

The aim of this project is to find an optimal location for the Award Ceremony.

The report will be targeted to the organizing committee of the Foundation Circle Halewijn, but could be useful to other stakeholders, such as book retailers and literary organizations.

## 2. DATA

In order to meet the above mentioned criteria, we should gather insight in

- \* existing neighborhoods of Haarlem, based on the Dutch Postal Code system;
- \* the number of existing bookshops in the various neighborhoods of Haarlem;
- \* the number of restaurants in the neighborhood, if any.

Delineation of neighborhoods will be based on the Dutch Postal Code dictionary.

### 2.1 Data sources

The following data sources were used to extract the information required:

\* [Nederlandse Postcodetabel](#) > Dutch Postal Code Dictionary in Excel xlsx format

NB: A [Dutch Postal Code API](#) is also available, but only as a paid service.

\* [GeoPy](#) > Python 3 client for geocoding web services: coordinates of Haarlem Railway Station.

\* [Foursquare API](#) > location data of book shops and restaurants.

## 2.2 Data cleaning and Feature selection

From the primary Dutch postal codes dataset `pd_Dutchpc` ( $n=471781$ ; 46.8Mb), all non-Haarlem records were dropped. From the resulting dataset ( $n=4257$ ) PO Box data (all with one geographical location) were discarded, leaving 4123 records.

Inspection of the dataframe revealed the presence of (10) missing values in the column `PostcodeLetters`. However, this column, as well as 7 other columns, were not informative with regard to current analysis, so they were dropped.

Then street names were aggregated into unique postal codes, but still 4032 entries remained.

Therefore, unique postal codes were further segmented into 19 different Neighborhoods (“Wijk”) (based on the Postal Code (“Postcodenummer”).

The features selected thus were: Neighborhood, Street Names, averaged latitudes and longitudes.

Coordinates of existing book shops and restaurants were added to the dataframe.

Inspection of the definite data revealed no missing values. All but one data formats were judged as adequate, i.e. datatype of column “Wijk” was changed from integer to object to enable labelling on the generated map.

After calculation of the number of book shops and number of restaurants per neighborhood, data were normalized to yield relative frequencies.

## 3. METHODOLOGY

In this project we will identify areas of Haarlem based on book shop density, particularly those with the presence of restaurants.

First, the city of Haarlem was split up in Postal Code neighborhoods. Maximum distance between any two nearby neighborhood centroids will be calculated.

Data are then collected on the number of book shops and restaurants for each neighborhood within a radius equivalent of half this maximum distance.

Relative frequencies of book stores and restaurants will be calculated for each neighborhood.

Neighborhoods are then clustered by means of K-Means clustering.

Folium was used to visually identify promising areas around the Railway Station with a high number of book stores and restaurants.

Eligible neighborhoods, based on the above mentioned criteria and taking into account the distance from the Haarlem Central Station, will be presented.

## 4. EXPLORATORY DATA ANALYSIS

### 4.1 Construction of the primary dataset

Primary postal code Excel file was downloaded as zip-file, extracted and stored locally. However, this file contained 471781 records! To spare computational resources, this file was stripped to 4257 Haarlem-only records in Excel<sup>R</sup> and stored as Haarlempostalcode.xlsx file.

	PostcodeID	PostCodePK	PostCode	PostcodeNummers	PostcodeLetters	Straat	MinNummer	MaxNummer	Plaats	Gemeente	Provincie	Latitude	Longitude
0	258150	2000AA_0	2000AA	2000	AA	Postbus	0	10000	Haarlem	Haarlem	Noord-Holland	52.381016	4.64567
1	430279	2000AB_0	2000AB	2000	AB	Postbus	0	10000	Haarlem	Haarlem	Noord-Holland	52.381016	4.64567
2	79515	2000AC_0	2000AC	2000	AC	Postbus	0	10000	Haarlem	Haarlem	Noord-Holland	52.381016	4.64567
3	88048	2000AD_0	2000AD	2000	AD	Postbus	0	10000	Haarlem	Haarlem	Noord-Holland	52.381016	4.64567
4	262603	2000AE_0	2000AE	2000	AE	Postbus	0	10000	Haarlem	Haarlem	Noord-Holland	52.381016	4.64567

**Table 1.** Raw dataset of Haarlem postal code areas with geodata (n = 4257).

### 4.2 Data cleaning and Feature selection

'Postbus' (PO Box) values in column Straat are not informative (all with same geolocation). They were dropped.

Inspection of this dataset revealed 10 missing values in the column PostcodeLetters. However, for our analysis we only need the columns PostCode, PostcodeNummers, Straat, Latitude and Longitude, so the other columns were dropped from analysis:

	PostCode	PostcodeNummers	Straat	Latitude	Longitude
0	2011AA	2011	Spaarnwouderstraat	52.379617	4.641031
1	2011AA	2011	Melkboersteeg	52.381016	4.645670
2	2011AB	2011	Sleutelstraat	52.380279	4.644150
3	2011AB	2011	Spaarnwouderstraat	52.380131	4.643988
4	2011AC	2011	Koralensteeg	52.380790	4.645376

**Table 2.** Adapted dataset 1 of Haarlem postal code areas with geodata (n = 4123).

As shown in Table 2, duplicates were present in the column for postal codes ("PostCode"). So, street names pertaining to one postal code were aggregated in a new column "StraatNamen". The average latitude and longitude for each postal code was then added to this dataframe:

	PostCode	PostcodeNummers	StraatNamen	Latitude_m	Longitude_m
	2011AA	2011	Spaarnwouderstraat, Melkboersteeg	52.380317	4.643351
	2011AB	2011	Sleutelstraat, Spaarnwouderstraat	52.380205	4.644069
	2011AC	2011	Koralensteeg, Spaarnwouderstraat	52.380699	4.645494
	2011AD	2011	Spaarnwouderstraat, Wijdesteeg	52.379602	4.641006
	2011AE	2011	Ossenhoofdsteeg, Wijdesteeg, Spaarnwouderstraat	52.379657	4.642028

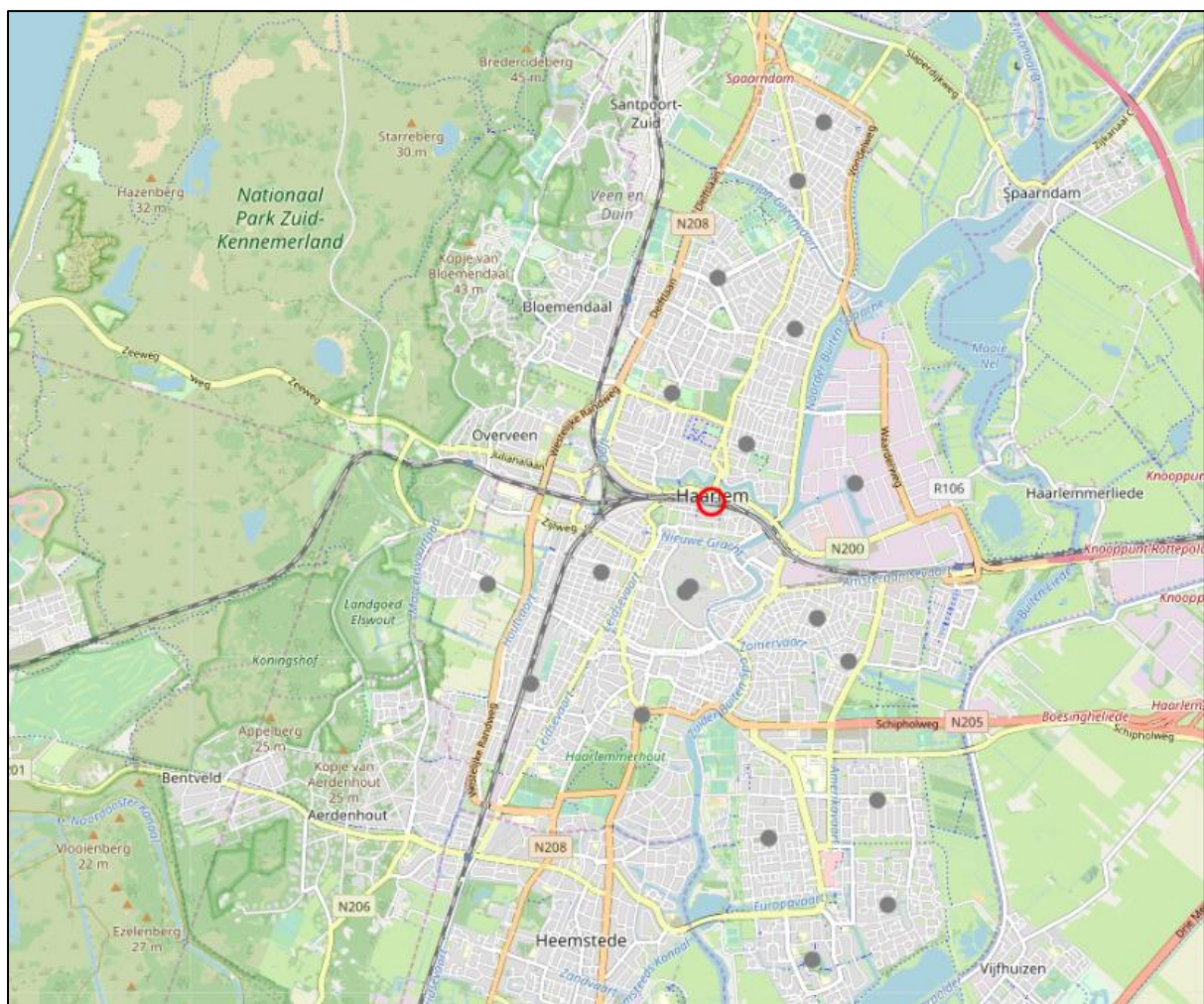
**Table 3.** Adapted dataset 2 of Haarlem postal code areas with averaged geodata; street names aggregated (n = 4032).

This dataset still contained a lot of records. Since 19 unique areas (“PostcodeNummers”) were identified, latitudes and longitudes for each unique area were (again) averaged; the column PostCode was dropped; and the column PostcodeNummers was renamed in Neighborhood (“Wijk”). Finally, data format for the column “Wijk” was changed from integer to object, so as to provide labels for each neighborhood in the generated map:

Wijk	StraatNamen	Latitude_m	Longitude_m
2011	Spaarnwouderstraat, Melkboersteeg, Sleutelstra...	52.381240	4.635808
2012	Zuider Buiten Spaarne, Zuider Buiten Spaarne, ...	52.370950	4.629562
2013	Duvenvoordestraat, Duvenvoordestraat, Duvenvoo...	52.382397	4.624165
2014	Willem Bontekoestraat, Houtvaartpad, Houtmanka...	52.373490	4.614884
2015	Tulpenkade, Krokusstraat, Krokusstraat, Krokus...	52.381513	4.609336

**Table 4.** Adapted dataset 3 of Haarlem postal code areas with geodata (n = 19).

#### 4.3 Building the primary map overview



**Figure 1.** Haarlem Map. Red open circle: Central Station location. Grey dots: neighborhood centroids (n = 19).


The top 5 distances between *nearby* neighborhood centroids were calculated as follows: 2012-2033: 1905 m, 2012-2032: 1792 m, 2019-2023: 1769 m, 2013-2023: 1717 m and between neighborhoods 2031 and 2022: 1472 m. Therefore, a radius of 1000 m was implemented in assigning features to a particular neighborhood (see 4.5).

Greatest distances between neighborhood centroids were calculated as follows: 2026-2036: 7459 m, 2026-2037: 6988 m, 2025-2036: 6943 m, 2025-2037: 6499 m, 2026-2034: 6390 m. Minimum distance between two postal code neighborhoods was 57 m (2011-2019).

#### 4.4 Creating definite map of Haarlem with feature (location) data

Name and Location of book stores and restaurants in the area were identified within a radius of 2500 m (i.e. in walking distance) from Central Station by means of Foursquare:

0	H. de Vries Boeken
1	Kennemer Boekhandel
2	Boekenvoordeel
3	Athenaeum Boekhandel
4	Boekenhoek
5	Plantage boekhandel
6	Boekhandel Bloemendaal
7	Gillissen & Co Boekhandel
8	't Boekendaaltje
9	Nautilus Boekbinderij



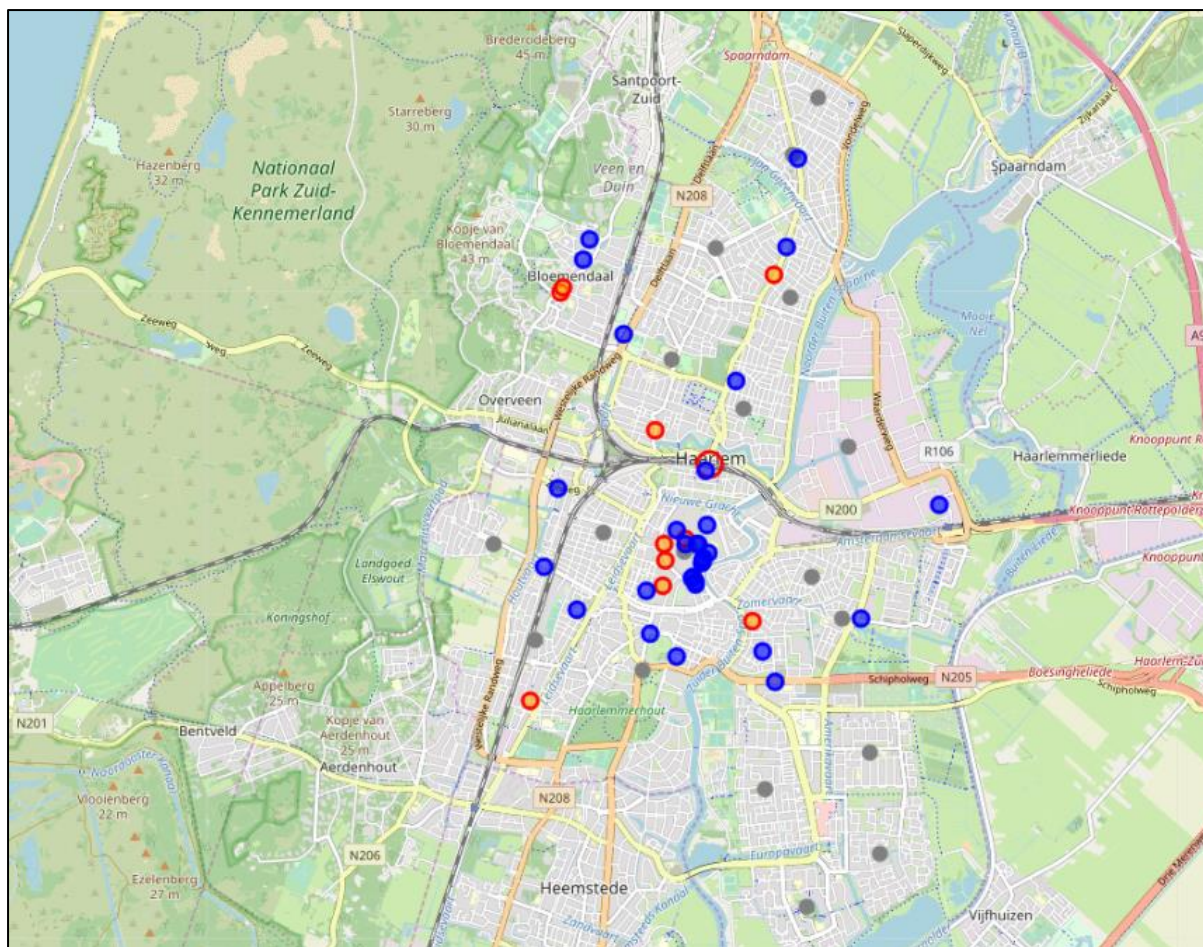
**Table 5.** Book stores present within radius of 2.5 km from Haarlem Central Station (n=10).

0	Café-Bar-Restaurant Beijneshal
1	Hotel-Restaurant Carillon
2	Restaurant Delphi
3	Restaurant ML
4	Restaurant Parck
5	Ma Browns Restaurant
6	Chinees Restaurant de Gouden Lelie
7	Restaurant Dubrovnik
8	Restaurant Ludic
9	Restaurant de Wandelaar
10	Restaurant De Generaal
11	Restaurant Applause
12	IKEA Restaurant
13	Restaurant Esperanto
14	Restaurant Ohm
15	Restaurant Het Pakhuis
16	Restaurant Scampi
17	Restaurant Babbels
18	restaurant van der Valk Haarlem
19	Restaurant Noor
20	Restaurant Sin Yue
21	Restaurant Kunstijsbaan Kennemerland
22	Restaurant Variee
23	Merdane Restaurant
24	Restaurant Fris
25	Restaurant Alkmaar
26	Restaurant Jeune & Agneau
27	Restaurant Noor
28	Restaurant NAP
29	Restaurant Reinaldahuis



**Table 6.** Restaurants present within radius of 2.5 km from Haarlem Central Station (n=30).





**Figure 2.** Haarlem Map. Red open circle: Central Station. Grey dots: neighborhood centroids ( $n = 19$ ). Orange dots: book store locations. Blue dots: restaurant locations.

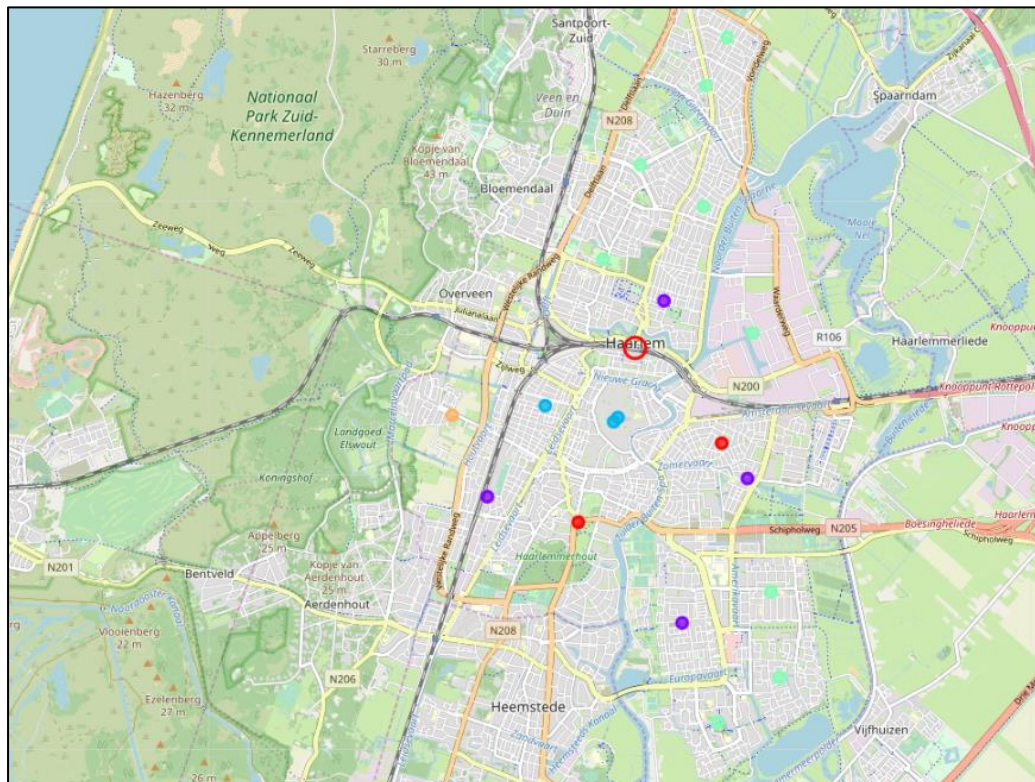
#### 4.5 Exploring 19 neighborhoods of Haarlem

By using Foursquare, book stores and restaurants, present within a radius of 1000 m from the neighborhood centroid (averaged latitude/longitude) were assigned to each neighborhood. For each feature relative frequencies were calculated as number of items in neighborhood / total number of items in the Haarlem area.

The 19 neighborhoods were then clustered on the basis of the relative frequencies of book stores and restaurants present:

Cluster	Wijk	StraatNamen	Latitude_m	Longitude_m	NrBook	RfBook	NrRest	RfRest
2	2011	Spaarnwouderstraat, Melkboersteeg, Sleutelstra...	52.381240	4.635808	5	0.161290	30	0.138249
0	2012	Zuider Buiten Spaarne, Zuider Buiten Spaarne, ...	52.370950	4.629562	2	0.064516	27	0.124424
2	2013	Duvenvoordestraat, Duvenvoordestraat, Duvenvoo...	52.382397	4.624165	5	0.161290	30	0.138249
1	2014	Willem Bontekoestraat, Houtvaartpad, Houtmanka...	52.373490	4.614884	2	0.064516	7	0.032258
4	2015	Tulpenkade, Krokusstraat, Krokusstraat, Krokus...	52.381513	4.609336	0	0.000000	10	0.046083
2	2019	Vroonhof, Achterlangs, Piet van Heerdenplein, ...	52.380887	4.635200	5	0.161290	30	0.138249
1	2021	Ben Viljoenstraat, Ben Viljoenstraat, Ben Vilj...	52.392686	4.643269	1	0.032258	8	0.036866
3	2022	Julianapark, Julianapark, Soendaplein, Floress...	52.401850	4.649644	1	0.032258	4	0.018433
3	2023	Schotersingel, Schotersingel, Schotersingel, S...	52.396750	4.633461	1	0.032258	4	0.018433
3	2024	Lodewijk van Deijssellaan, Lodewijk van Deijss...	52.405990	4.639572	1	0.032258	4	0.018433
3	2025	Vondelweg, Vondelweg, Vondelweg, Vondelweg, Vo...	52.413763	4.649940	1	0.032258	3	0.013825
3	2026	Leeuwendalersstraat, Leeuwendalersstraat, Jan ...	52.418401	4.653400	1	0.032258	1	0.004608
3	2031	Oude Waarderweg, Jacques Meuwissenweg, Maus Ga...	52.389543	4.657597	0	0.000000	1	0.004608
0	2032	Amsterdamsevaart, Amsterdamsevaart, Amsterdams...	52.378746	4.652589	1	0.032258	30	0.138249
1	2033	Robertus Nurksweg, Amsterdamsevaart, Liewegje,...	52.375255	4.656638	1	0.032258	7	0.032258
1	2034	Denemarkenstraat, Denemarkenstraat, Denemarken...	52.361146	4.646108	1	0.032258	8	0.036866
3	2035	Van Deventerstraat, Van Deventerstraat, Van De...	52.364156	4.660410	1	0.032258	5	0.023041
3	2036	Christine Koetsstraat, Christine Koetsstraat, ...	52.351372	4.651814	1	0.032258	5	0.023041
3	2037	Ceylonpoort, Mentonpassage, Rivi+gradreef, Riv...	52.355810	4.661876	1	0.032258	3	0.013825


**Table 7.** Cluster allocation based on relative frequencies of book stores and restaurants present (n=19).




**Figure 3.** Haarlem Map, neighborhoods clustered. Red open circle: Central Station. Clusters (n=19): #1 red dots (n=2); #2 purple (n=4); #3 blue (n=3); #4 green (n=9); #5 yellow (n=1).

#### 4.6 Cluster examination


Inspection of the various clusters:

Cluster nr 1 					
	Wijk	NrBook	NrRest	RFBook	RFRest
1	2012	2	27	0.064516	0.124424
13	2032	1	30	0.032258	0.138249


  

Cluster nr 2 					
	Wijk	NrBook	NrRest	RFBook	RFRest
3	2014	2	7	0.064516	0.032258
6	2021	1	8	0.032258	0.036866
14	2033	1	7	0.032258	0.032258
15	2034	1	8	0.032258	0.036866


  

Cluster nr 3 					
	Wijk	NrBook	NrRest	RFBook	RFRest
0	2011	5	30	0.16129	0.138249
2	2013	5	30	0.16129	0.138249
5	2019	5	30	0.16129	0.138249

Cluster nr 4 					
	Wijk	NrBook	NrRest	RFBook	RFRest
7	2022	1	4	0.032258	0.018433
8	2023	1	4	0.032258	0.018433
9	2024	1	4	0.032258	0.018433
10	2025	1	3	0.032258	0.013825
11	2026	1	1	0.032258	0.004608
12	2031	0	1	0.000000	0.004608
16	2035	1	5	0.032258	0.023041
17	2036	1	5	0.032258	0.023041
18	2037	1	3	0.032258	0.013825

Cluster nr 5 					
	Wijk	NrBook	NrRest	RFBook	RFRest
4	2015	0	10	0.0	0.046083

**Table 8.** Cluster parameters: number (Nr) and relative frequencies (RF) of book stores and restaurants present.



From table 8 it might be concluded, that

- \* **cluster 1** contains two neighborhoods with only a few book stores; in contrast, lots of restaurants are present. Calculated distance from Central Station are 2002 m [2012] and 1410 m [2032].
- \* **cluster 2** comprises of 4 neighborhoods with eccentric location and relatively large distance from the Station. Over here, one only finds a limited number of book stores and restaurants;
- \* **cluster 3** encompass 3 nearby neighborhoods (2011-2013-2019), centrally located, with a high number of book stores as well as restaurants. The average distance from Central Station is only 927 meters (2011:CS = 783 m, 2013:CS = 1167 m, 2019:CS = 832 m), i.e. a 10-minute walk;
- \* **cluster 4** contains the largest number of neighborhoods (n=9), with obvious eccentric location. The average distance from Haarlem Central Station is 2630 meters, i.e. a >30-minutes walk (min 1025 m [2023], max 4182 m [2036]);
- \* **cluster 5** consist of only one neighborhood [2015] with no book stores present but an average number of restaurants.

## 5. RESULTS AND DISCUSSION

The city of Haarlem can be segmented into 19 neighborhoods, based on the available Dutch Postal Code system (fig.1). Lots of book stores as well as restaurants are present, however, not evenly distributed. From fig.2 it can be seen that most book stores are concentrated in the center of Haarlem, i.e. within walking distance of the Central Station. Restaurants on the other hand, are widely distributed in the whole area.

The challenge was to find an optimum spot for the yearly Halewijn Award Ceremony. The criteria formulated by the Organizing Committee were as follows: a) a high concentration of book stores and b) many restaurants present in the neighborhood; preferably within walking distance from the Central Station (20 minutes = 2 kilometers).

Based on these criteria, five clusters of neighborhoods were identified (fig.3, table 8).

Cluster 3 encompass three nearby neighborhoods with a high number of book stores as well as restaurants. The average distance from Central Station is only 927 meters (2011:CS = 783 m, 2013:CS = 1167 m, 2019:CS = 832 m).

Cluster 1 contains two neighborhoods with only a few book stores but a large number of restaurants and still within walking distance of the Central Station.

Holding the yearly ceremony in one of the neighborhoods in cluster 3 seems obvious, since it combines large numbers of book stores and restaurants. However, one might consider other places located in cluster 1, especially if priority is given to a less crowded area for the dining afterwards.

## 6. CONCLUSION

From the results so far, we might conclude that optimal neighborhoods for the Halewijn Award Ceremony comprise postal code areas 2011, 2013 and 2019 with high numbers of book stores and restaurants present, within walking distance of the Central Railway Station. Further exploration of neighborhoods 2012 and 2032 however is encouraged, especially if dining after the ceremony in a quiet surrounding is desired.

On spot visualization of these neighborhoods is suggested, in order to make a reasonable choice for the forthcoming location of the Halewijn Award ceremony in Haarlem.