Table of Contents

1 Introduction

1.1 Background

People with Asthma can benefit to live in an area with has a low number PM2.5.
PM2.5 refers to atmospheric particulate matter (PM) that have a diameter of less than 2.5 micrometers, which is about 3% the diameter of a human hair.
Exposure to such particles can affect both your lungs and your heart. Numerous scientific studies have linked particle pollution exposure to a variety of problems.

1.2 Problem

In the Netherlands you can use a website from the government to see PM numbers and also other Air Pollution measurements.
It does not show things like Hospitals, Schools, House prices etc.
So, if you want to find a good area to live you need to look into different sources and try to relate this information to each other.

If you can show this kind of information on a geographical map together this will make it much clearer at one glance what could be possible good areas to live in.

One thing I was interested in some 8 years ago was if there was a hospital nearby places with low pollution. I was not interested in other topics at the time. I could not find all information in one place.

1.3 Interest

People who want to live in an area with low pollution, and at the same time want to know what other things are of interest in the area.
This will be different for everyone. You can think of Hospitals, House Prices, Shopping Centers, Schools, Unemployment rate etc.

In order not to make things too complex in this assignment I will focus on the following:
- finding the least polluted area with hospitals within 7 km.

The program can be extended with other features later if needed.

2 Data Acquisition and Cleaning

2.1 Data sources

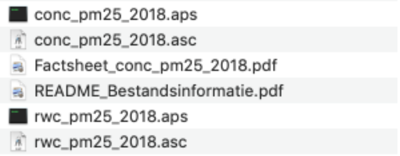
2.1.1 PM2.5 data from RIVM

RIVM is the National Institute for Public Health and Environment in the Netherlands.

The dataset for PM25 of the year 2018 is downloaded from https://geodata.rivm.nl/gcn/ in a ASCII-GRID format. See also https://en.wikipedia.org/wiki/Esri_grid

Locations in the grid are according to the 'Amersfoortse coordinaten' which is described in dutch on https://nl.wikipedia.org/wiki/Rijksdriehoeksco%C3%B6rdinatenAmersfoort has a location of X=155000 and Y=463000

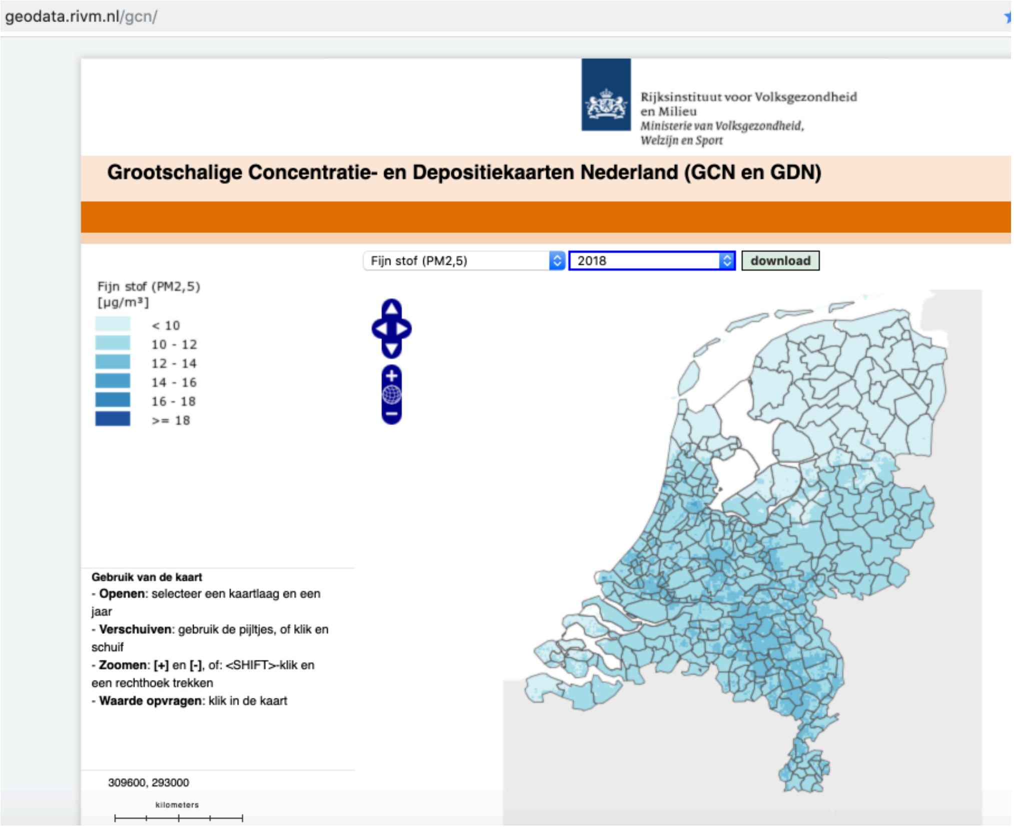
Downloading gives you a zip file with the following files



conc_pm25_2018.aps
conc_pm25_2018.asc
Factsheet_conc_pm25_2018.pdf
README_Bestandsinformatie.pdf
rwc_pm25_2018.aps
rwc_pm25_2018.asc

It is stored in folder con_pm25_2018.

The file conc_pm25_2018.asc contains the data I used in the assignment.
Both .aps and .asc contains same information but in a different format.
The files starting with rwc contains additional information on highways which I do not need for the assignment.



geodata.rivm.nl/gcn/

Rijksinstituut voor Volksgezondheid en Milieu
Ministerie van Volksgezondheid, Welzijn en Sport

Grootschalige Concentratie- en Depositiekaarten Nederland (GCN en GDN)

Fijn stof (PM2,5) | 2018 | download

Fijn stof (PM2,5)
[μg/m³]

- ■ < 10
- ■ 10 - 12
- ■ 12 - 14
- ■ 14 - 16
- ■ 16 - 18
- ■ >= 18

Gebruik van de kaart
- **Openen**: selecteer een kaartlaag en een jaar
- **Verschuiven**: gebruik de pijltjes, of klik en schuif
- **Zoomen**: [+] en [-], of: <SHIFT>-klik en een rechthoek trekken
- **Waarde opvragen**: klik in de kaart

309600, 293000

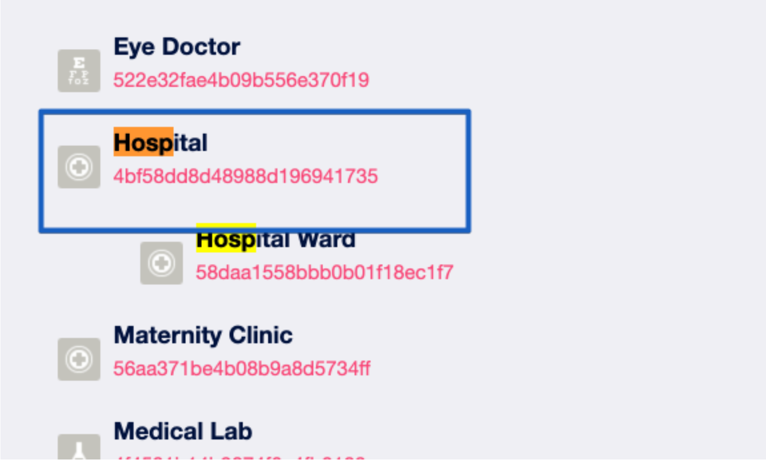
Figure 1 From this site Downloaded PM25 data for 2018

2.1.2 Hospitals using Foursquare

I use the search parameter since I am only interested in locations and not in reviews.

```
url = 'https://api.foursquare.com/v2/venues/search?client_id={}&client_secret={}&ll={},{}&v={}&radius={}&limit={}&categoryId={}'\
.format(CLIENT_ID, CLIENT_SECRET, latitude, longitude, VERSION, radius, LIMIT,cat)
```

For category I use 4bf58dd8d48988d196941735 as described on Foursquare https://developer.foursquare.com/docs/resources/categories



Eye Doctor
522e32fae4b09b556e370f19

Hospital
4bf58dd8d48988d196941735

Hospital Ward
58daa1558bbb0b01f18ec1f7

Maternity Clinic
56aa371be4b08b9a8d5734ff

Medical Lab

2.2 Data Cleaning

2.2.1 PM25 Reduce number cells

The PM25 data has a shape of 320*280 = 89600 cells. Each cell represents 1 km x 1 km.

89600 is too much to handle using Nominatim to get address details. Nominatim limits the number of calls to max 1 call per second. I set the sleep timer to 2 seconds to be sure I am within limits.

A cell size of 3 km x 3 km which still gives good results for PM25 data. Thereby reducing the number of cells to 9955.
The data contains a lot of 'null' values represented by -999.0. This is mainly data representing the North Sea.
If you have a grid which needs to cover the Netherlands a lot of water is included.
See below

Leeuwarden
Groningen
E22
E232
Alkmaar
E22
Amsterdam
E232
E35
E30
The Hague
Netherlands
Arnhem
Rotterdam
E31
E311
E25
3
57
Eindhoven
Essen
E19
Antwerp
E34
Düsseldorf
D

Replacing -999.0 by NaN and use filtering the number of cells is reduced to 4886.
Getting address details will now use approximately 162 minutes.

2.2.2 PM25 location conversion to latitude/longitude

Location in the PM25 grid is according to the 'Amersfoortse coordinaten' which is described in Dutch on https://nl.wikipedia.org/wiki/Rijksdriehoeksco%C3%B6rdinatenAmersfoort has a location of X=155000 and Y=463000

First 6 lines show

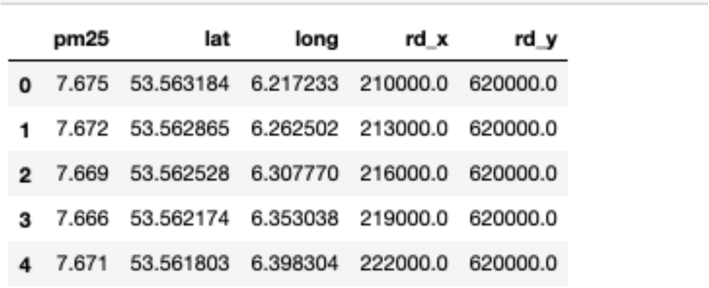
- NCOLS 280 (nr of cells in X direction)
- NROWS 320 (nr of cells in Y direction)
- XLLCORNER 0 (X coordinate)
- YLLCORNER 300000 Y Coordinate
- CELLSIZE 1000 (1000 meters so 1 km)

- NODATA_VALUE -0.9990E+03 (this is the value if there is no data)

To convert the data into longitude and latitude I use code from https://thomasv.nl/2014/03/rd-naar-gps/ This converts the so called Rijksdriehoek coordinates(rd_x and rd_y) into latitude and longitude. Conversion code is a separate class called RDWGSConverter

The Netherlands is 320km height by 280 width.
The number of rows in PM25 data 320. A row represents the Y coordinate.
A row number starts with 0 and the increments
The Y coordinate starts at 620 and decrements.
So, row 0 is y-coordinate 620

Coordinates are represented in meters. So, I have to multiply a grid position by 1000 meters since a cell is 1 km.
The result is shown below

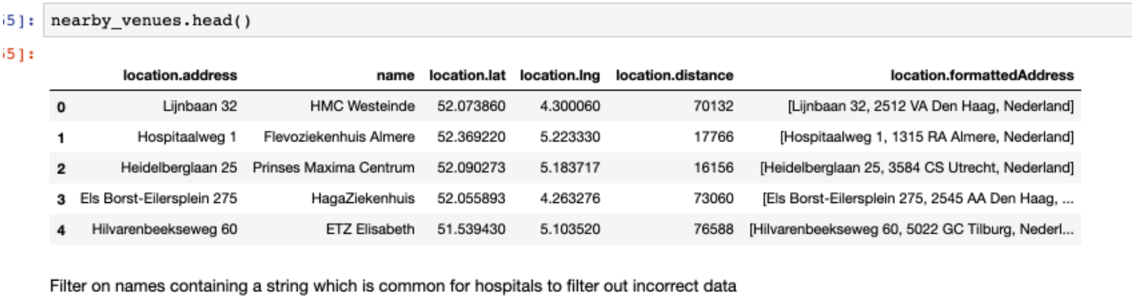


	pm25	lat	long	rd_x	rd_y
0	7.675	53.563184	6.217233	210000.0	620000.0
1	7.672	53.562865	6.262502	213000.0	620000.0
2	7.669	53.562528	6.307770	216000.0	620000.0
3	7.666	53.562174	6.353038	219000.0	620000.0
4	7.671	53.561803	6.398304	222000.0	620000.0

2.2.3 Hospitals

When using the radius of 172 KM and central location of Baarn no cleaning is needed.
The list looks good. Looking into the exported file to CSV visually all locations look valid.

If I would use a different kind of radius the result is different. Then I would do a filtering based on the name containing ‘ziekenhuis’ which is Dutch for Hospital and MC for Medical Centre. Unfiltered data see below

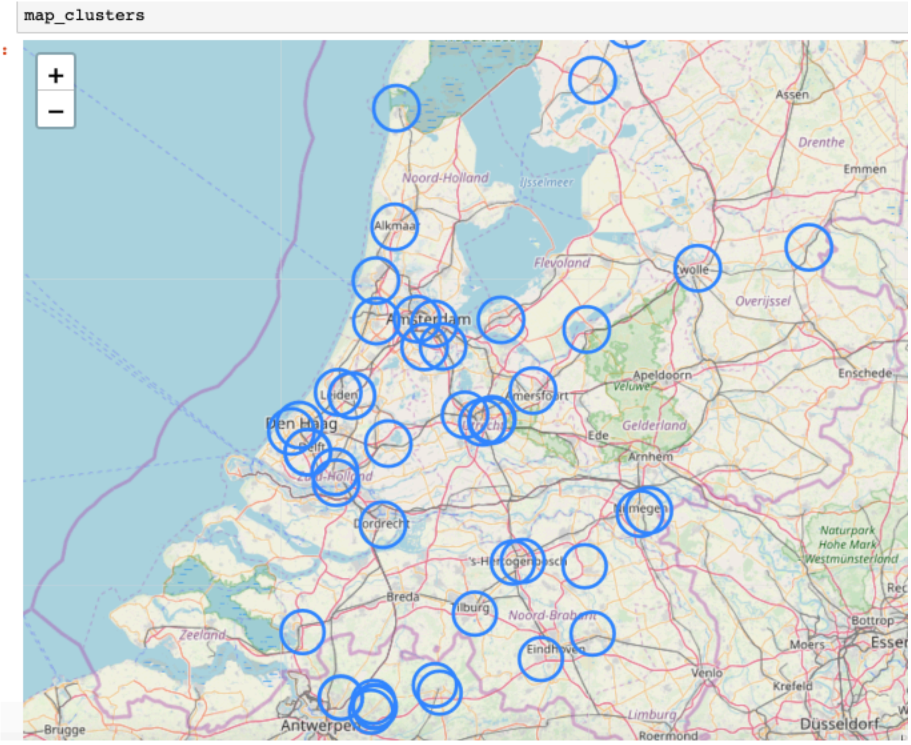


```
5]: nearby_venues.head()
```

5]:

	location.address	name	location.lat	location.lng	location.distance	location.formattedAddress
0	Lijnbaan 32	HMC Westeinde	52.073860	4.300060	70132	[Lijnbaan 32, 2512 VA Den Haag, Nederland]
1	Hospitaalweg 1	Flevoziekenhuis Almere	52.369220	5.223330	17766	[Hospitaalweg 1, 1315 RA Almere, Nederland]
2	Heidelberglaan 25	Prinses Maxima Centrum	52.090273	5.183717	16156	[Heidelberglaan 25, 3584 CS Utrecht, Nederland]
3	Els Borst-Eilersplein 275	HagaZiekenhuis	52.055893	4.263276	73060	[Els Borst-Eilersplein 275, 2545 AA Den Haag, ...
4	Hilvarenbeekseweg 60	ETZ Elisabeth	51.539430	5.103520	76588	[Hilvarenbeekseweg 60, 5022 GC Tilburg, Nederl...

Filter on names containing a string which is common for hospitals to filter out incorrect data



```
map_clusters
```

3 Methodology

3.1 Exploratory data analysis

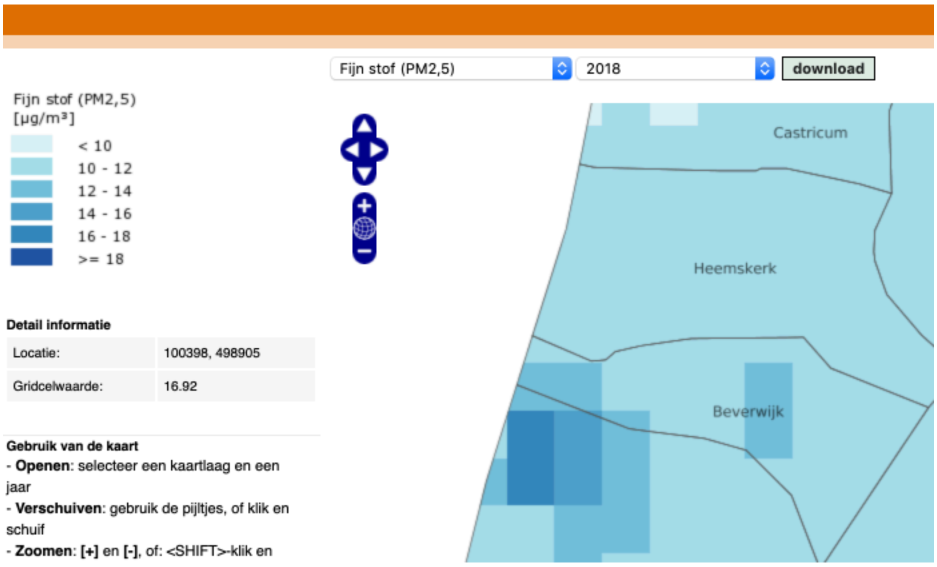
3.1.1 Get Geolocation data for PM25

To be able to convert a grid representing the Netherlands into latitude and longitude I needed to figure out how to read the data first.

The Netherlands is 320km height by 280 width.
The number of rows in PM25 data 320. A row represents the Y coordinate.
A row number starts with 0 and the increments
The Y coordinate starts at 620 and decrements.
So, row 0 is y-coordinate 620

Looking to how the coordinates where represented on the RIVM map and how this was represented in the rows and columns of the grid I was able to determine if my conversion from row and column number into a RijksCoordinate was correct (rd_x,rd_y).

I looked for example on the map from RIVM to find a location with high pollution.
There was only one cell in the data containing this number so I could confirm my conversion was correct



Fijn stof (PM2,5)
[µg/m³]

	< 10
	10 - 12
	12 - 14
	14 - 16
	16 - 18
	>= 18

Detail informatie

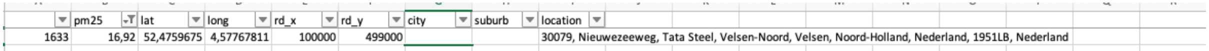
Locatie:	100398, 498905
Gridcelwaarde:	16.92

Gebruik van de kaart
- **Openen**: selecteer een kaartlaag en een jaar
- **Verschuiven**: gebruik de pijltjes, of klik en schuif
- **Zoomen**: [+] en [-], of: <SHIFT>-klik en

Figure 2 PM value is 16.92

The location is around Tata Steel where they produce steel.
From this map on RIVM website (see also earlier in this report) the location is at 100398, 498905

Looking into the excel generated from the PM25 dataframe



	pm25	lat	long	rd_x	rd_y	city	suburb	location
1633	16,92	52,4759675	4,57767811	100000	499000			30079, Nieuwezeeweg, Tata Steel, Velsen-Noord, Velsen, Noord-Holland, Nederland, 1951LB, Nederland

I could see only van PM25 value of 16,92 where I determined the grid cell location to be at 10000,499000 which is correct as the values are rounded on 1000 meter.

3.1.2 Get Hospital locations using Foursquare

Initially I wanted to find Hospitals near 'clean' locations so wanted to call Foursquare for every location.
Then I realized I would have a lot of duplicate hospitals with different distances to several locations which I would have to clean up.
Actually, the distance would not be relevant in my later calculations as I want to visualize clean locations on a map combined with location of hospitals.
When exploring how to map the data I realized I could plot a hospital location with a circle of radius 7 KM. So, on the map it would be visible where the clean locations are and also if this area would be within 7 km range of a hospital.

I noticed Foursquare returns different results based on the radius. Using a large radius, the data looks good.

See below the difference in using radius.
If I use Location Baarn with a radius of 172 KM I get 50 Hospitals which all look like real hospitals. Around Den Helder there is only 1 location/Hospital which is correct.
I know this since I live there.

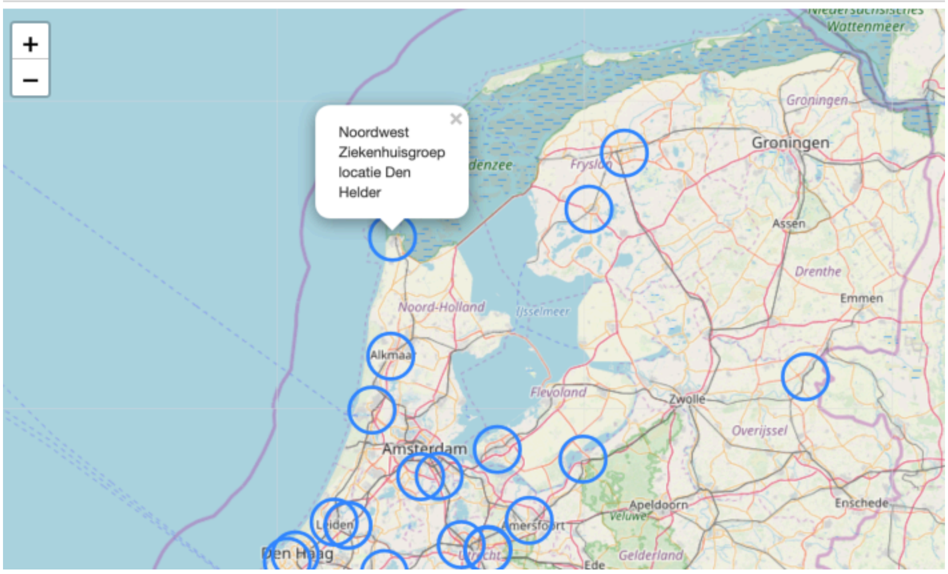
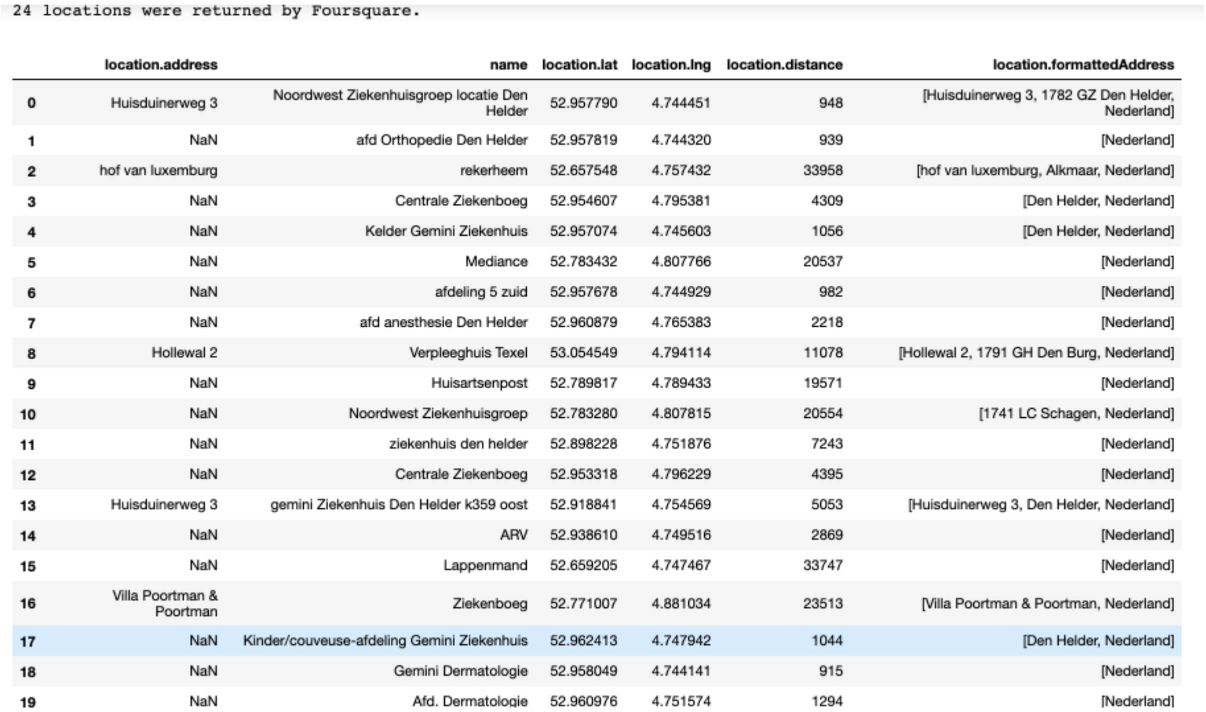


Figure 3 1 hospital near Den Helder

Distance between Den Helder and Alkmaar is some 30 km

If I use Location Den Helder with radius 30 KM I get 24 locations of which a lot of them are not hospitals.



```
24 locations were returned by Foursquare.
```

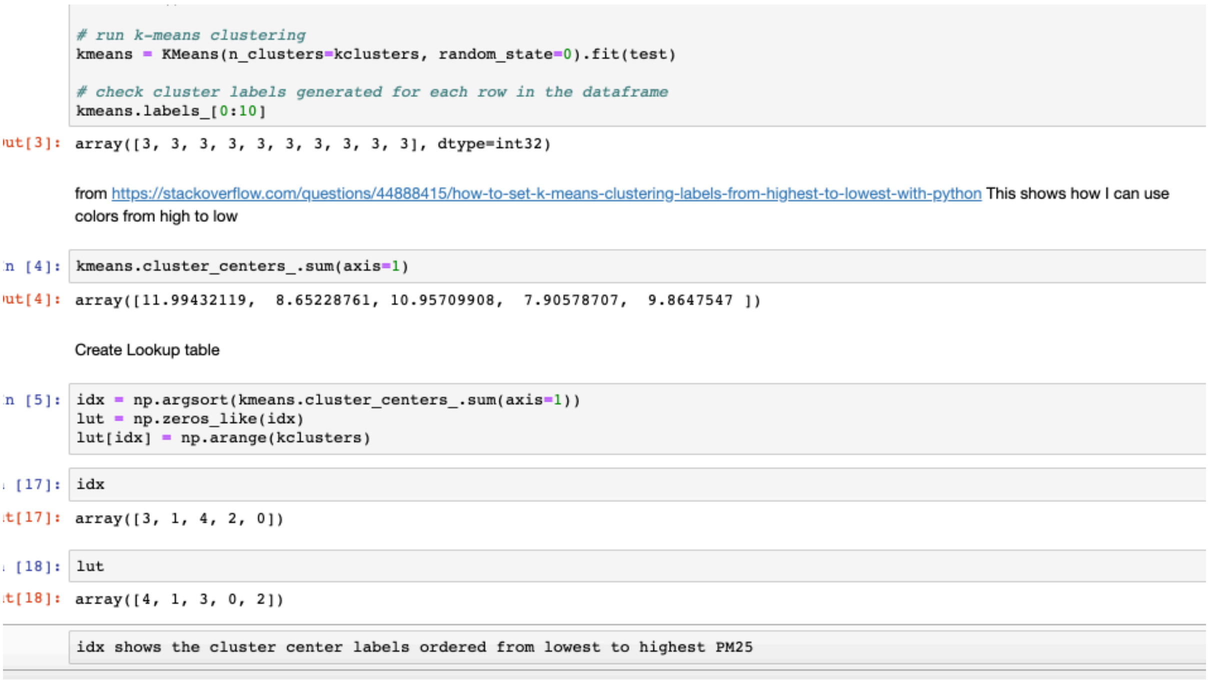
	location.address	name	location.lat	location.lng	location.distance	location.formattedAddress
0	Huisduinerweg 3	Noordwest Ziekenhuisgroep locatie Den Helder	52.957790	4.744451	948	[Huisduinerweg 3, 1782 GZ Den Helder, Nederland]
1	NaN	afd Orthopedie Den Helder	52.957819	4.744320	939	[Nederland]
2	hof van luxemburg	rekerheem	52.657548	4.757432	33958	[hof van luxemburg, Alkmaar, Nederland]
3	NaN	Centrale Ziekenboeg	52.954607	4.795381	4309	[Den Helder, Nederland]
4	NaN	Kelder Gemini Ziekenhuis	52.957074	4.745603	1056	[Den Helder, Nederland]
5	NaN	Mediance	52.783432	4.807766	20537	[Nederland]
6	NaN	afdeling 5 zuid	52.957678	4.744929	982	[Nederland]
7	NaN	afd anesthesie Den Helder	52.960879	4.765383	2218	[Nederland]
8	Hollewal 2	Verpleeghuis Texel	53.054549	4.794114	11078	[Hollewal 2, 1791 GH Den Burg, Nederland]
9	NaN	Huisartsenpost	52.789817	4.789433	19571	[Nederland]
10	NaN	Noordwest Ziekenhuisgroep	52.783280	4.807815	20554	[1741 LC Schagen, Nederland]
11	NaN	ziekenhuis den helder	52.898228	4.751876	7243	[Nederland]
12	NaN	Centrale Ziekenboeg	52.953318	4.796229	4395	[Nederland]
13	Huisduinerweg 3	gemini Ziekenhuis Den Helder k359 oost	52.918841	4.754569	5053	[Huisduinerweg 3, Den Helder, Nederland]
14	NaN	ARV	52.938610	4.749516	2869	[Nederland]
15	NaN	Lappenmand	52.659205	4.747467	33747	[Nederland]
16	Villa Poortman & Poortman	Ziekenboeg	52.771007	4.881034	23513	[Villa Poortman & Poortman, Nederland]
17	NaN	Kinder/couveuse-afdeling Gemini Ziekenhuis	52.962413	4.747942	1044	[Den Helder, Nederland]
18	NaN	Gemini Dermatologie	52.958049	4.744141	915	[Nederland]
19	NaN	Afd. Dermatologie	52.960976	4.751574	1294	[Nederland]

Figure 4 Den Helder and radius 30 km

So, based on above results I cannot rely on Foursquare to get the right results.
But for this assignment it is good enough.

3.2 K-Means to group into clusters ranging from low to high

I divide the pollution data ranging from low to high in 5 buckets.
In this assignment only 1 column is used which is PM25 values.
So, what happens it will have 5 buckets representing the average of pollution and then all the locations which are around this average.
Since K-Means is not sorted I have to sort it to get a range from low to high
I used https://stackoverflow.com/questions/44888415/how-to-set-k-means-clustering-labels-from-highest-to-lowest-with-python as the basis for my code.



```
# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(test)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

```
ut[3]: array([3, 3, 3, 3, 3, 3, 3, 3, 3, 3], dtype=int32)
```

from https://stackoverflow.com/questions/44888415/how-to-set-k-means-clustering-labels-from-highest-to-lowest-with-python This shows how I can use colors from high to low

```
n [4]: kmeans.cluster_centers_.sum(axis=1)
ut[4]: array([11.99432119,  8.65228761, 10.95709908,  7.90578707,  9.8647547 ])
```

Create Lookup table

```
n [5]: idx = np.argsort(kmeans.cluster_centers_.sum(axis=1))
       lut = np.zeros_like(idx)
       lut[idx] = np.arange(kclusters)
```

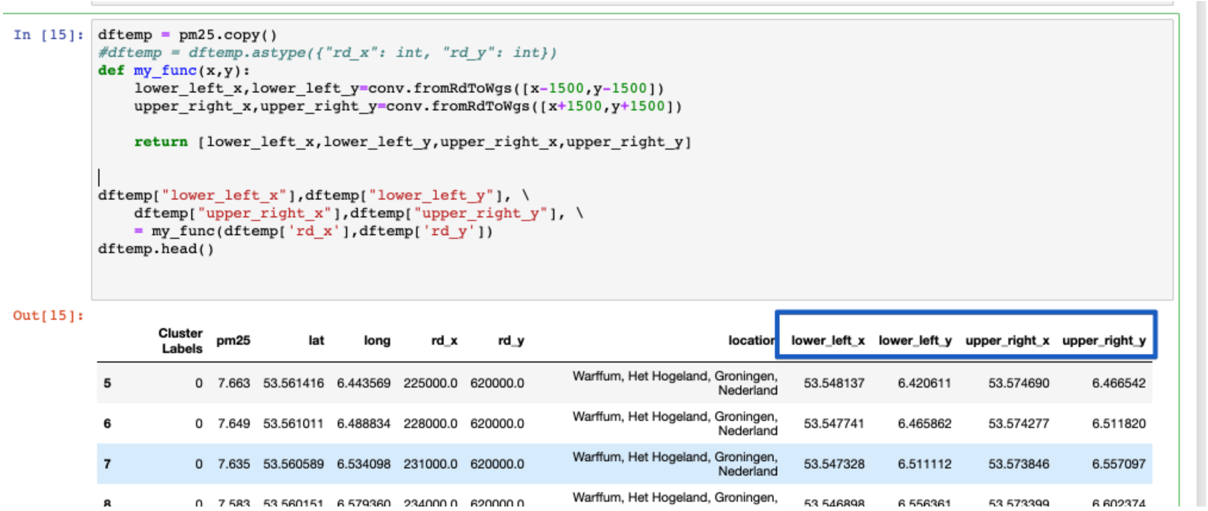
```
 [17]: idx
t[17]: array([3, 1, 4, 2, 0])
```

```
 [18]: lut
t[18]: array([4, 1, 3, 0, 2])
```

```
idx shows the cluster center labels ordered from lowest to highest PM25
```

3.3 Map to rectangles

To be able to represent the PM grid locations a rectangles of 3x3 km I need to add coordinates to reflect lower left and upper right corner.
To do this I needed to use the location and then go down/left 1500 meters
And up/right 1500 meters.
I did this based on the ‘rijkscoordinates’ which were then converted to proper latitude and longitude.



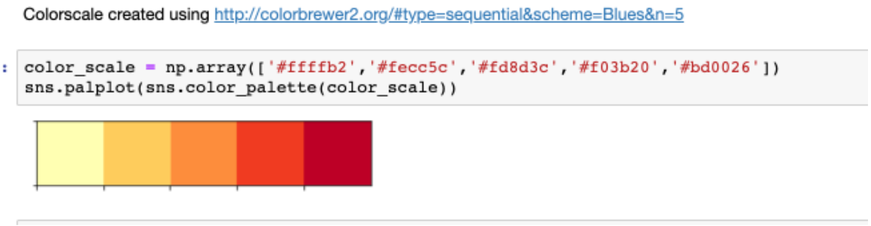
```
In [15]: dftemp = pm25.copy()
         #dftemp = dftemp.astype({"rd_x": int, "rd_y": int})
         def my_func(x,y):
             lower_left_x,lower_left_y=conv.fromRdToWgs([x-1500,y-1500])
             upper_right_x,upper_right_y=conv.fromRdToWgs([x+1500,y+1500])

             return [lower_left_x,lower_left_y,upper_right_x,upper_right_y]

         |
         dftemp["lower_left_x"],dftemp["lower_left_y"], \
             dftemp["upper_right_x"],dftemp["upper_right_y"], \
             = my_func(dftemp['rd_x'],dftemp['rd_y'])
         dftemp.head()
```

Out[15]:

	Cluster Labels	pm25	lat	long	rd_x	rd_y	locatior	lower_left_x	lower_left_y	upper_right_x	upper_right_y
5	0	7.663	53.561416	6.443569	225000.0	620000.0	Warffum, Het Hogeland, Groningen, Nederland	53.548137	6.420611	53.574690	6.466542
6	0	7.649	53.561011	6.488834	228000.0	620000.0	Warffum, Het Hogeland, Groningen, Nederland	53.547741	6.465862	53.574277	6.511820
7	0	7.635	53.560589	6.534098	231000.0	620000.0	Warffum, Het Hogeland, Groningen, Nederland	53.547328	6.511112	53.573846	6.557097
8	0	7.583	53.560151	6.579360	234000.0	620000.0	Warffum, Het Hogeland, Groningen,	53.546898	6.556361	53.573399	6.602374

4 Color Scale

To reflect the PM25 data from low to high I used a color scale.



Colorscale created using http://colorbrewer2.org/#type=sequential&scheme=Blues&n=5

```
: color_scale = np.array(['#ffffb2','#fecc5c','#fd8d3c','#f03b20','#bd0026'])
  sns.palplot(sns.color_palette(color_scale))
```

So yellow is low, red is high.

5 Results

I have created a map for Netherlands as a whole and a separate one for North-Holland. The map is similar to RIVM maps from the government but contains also the hospitals on the map.

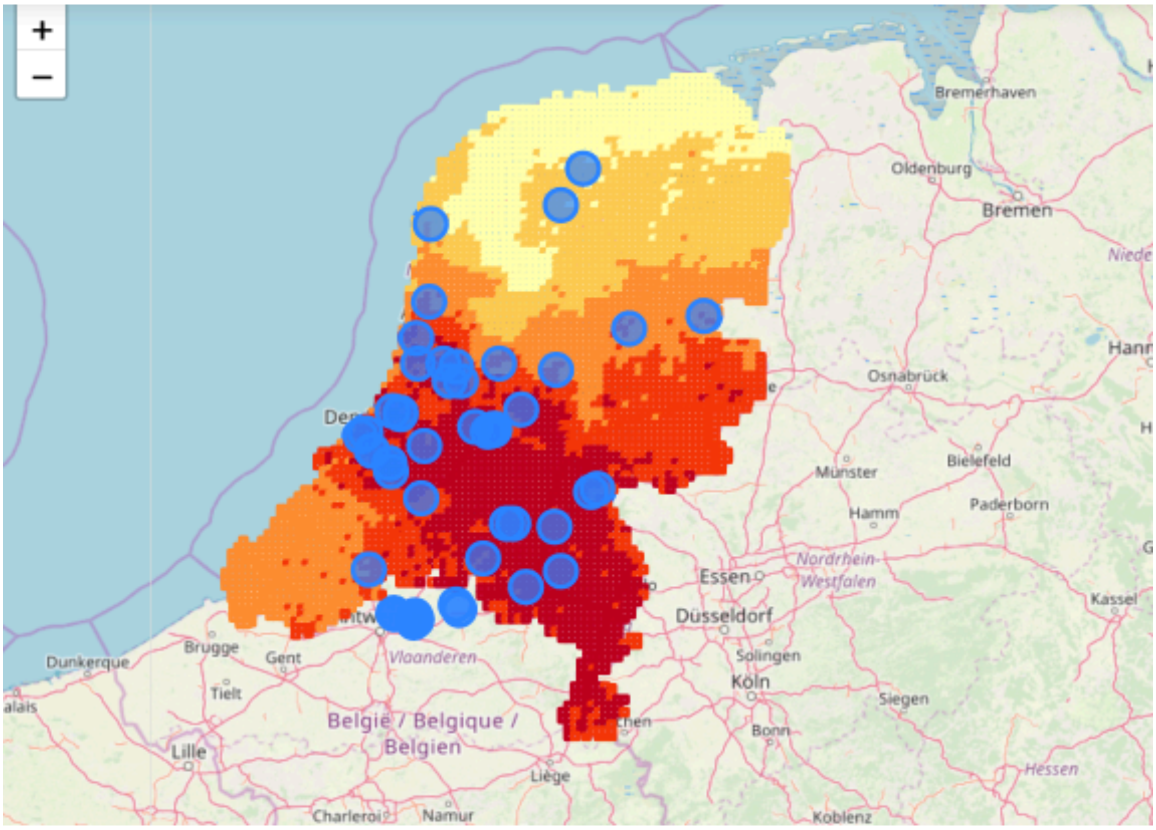
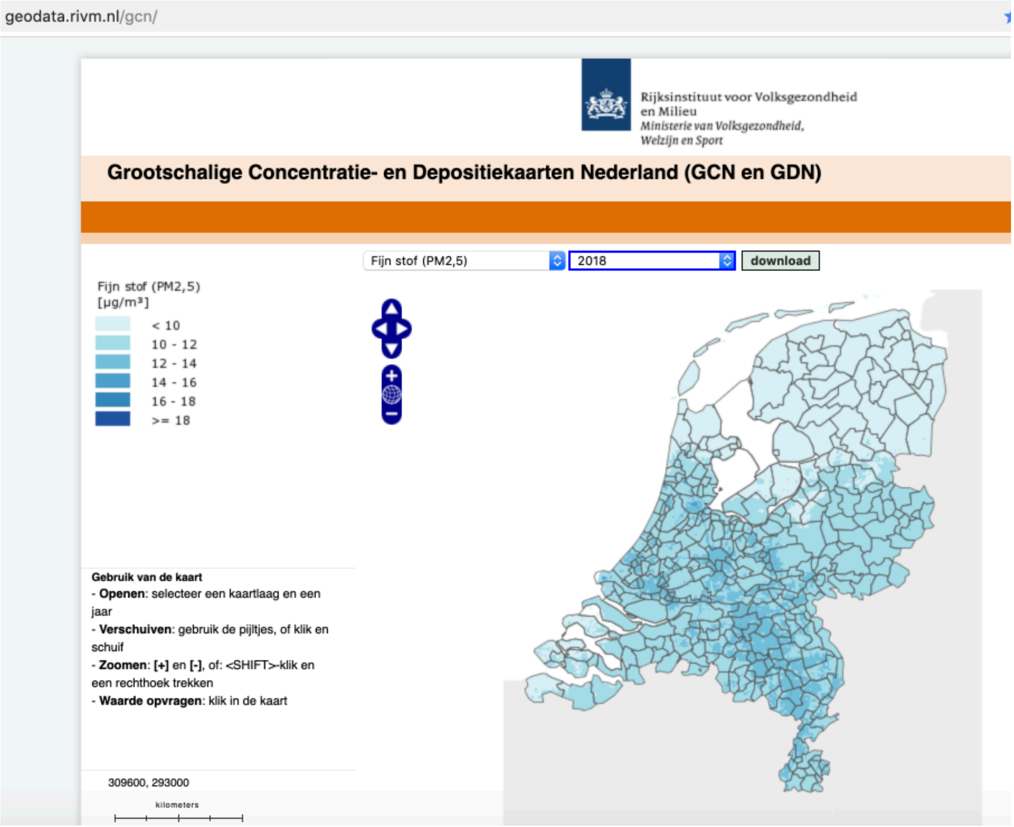


Figure 5 Netherlands PM25 and Hospitals

The map of the Netherlands show that the northern part of the Netherlands has lower pollution then the rest. The blue circles are hospitals.
Below is the map from RIVM which is based on 1by 1km. The chart above which I generated is 3x3 km. But they are pretty similar.



geodata.rivm.nl/gcn/

Rijksinstituut voor Volksgezondheid en Milieu
Ministerie van Volksgezondheid, Welzijn en Sport

Grootschalige Concentratie- en Depositiekaarten Nederland (GCN en GDN)

Fijn stof (PM2,5) | 2018 | download

Fijn stof (PM2,5)
[µg/m³]

- < 10
- 10 - 12
- 12 - 14
- 14 - 16
- 16 - 18
- >= 18

Gebruik van de kaart
- **Openen**: selecteer een kaartlaag en een jaar
- **Verschuiven**: gebruik de pijltjes, of klik en schuif
- **Zoomen**: [+] en [-], of: <SHIFT>-klik en een rechthoek trekken
- **Waarde opvragen**: klik in de kaart

309600, 293000

Figure 6 Original map from RIVM

Since I had a preference to live in North-Holland I decided to have a closer look there. See below.

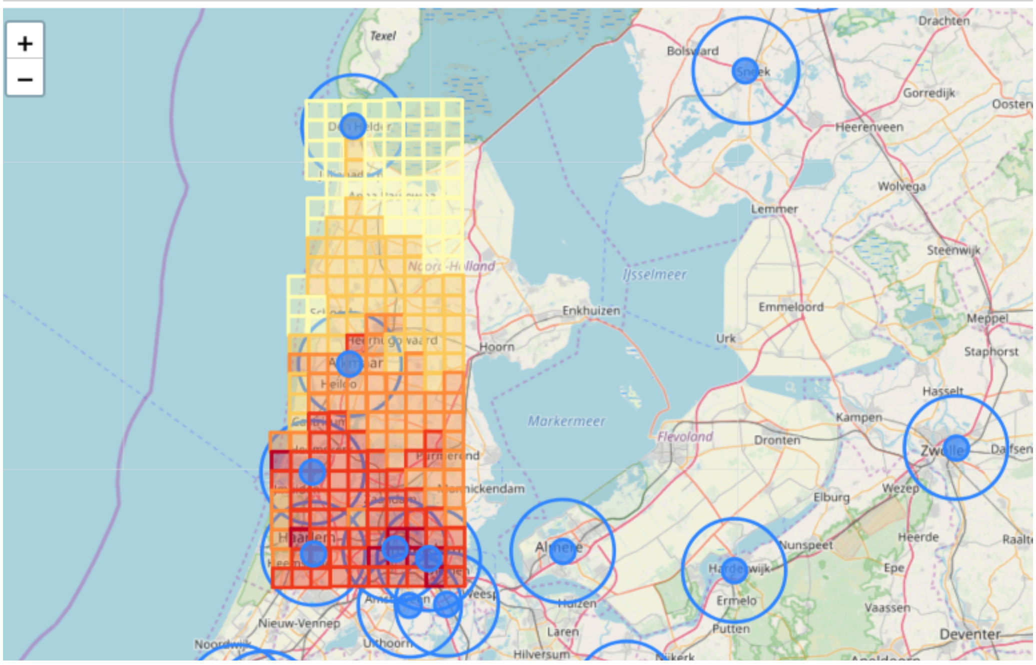
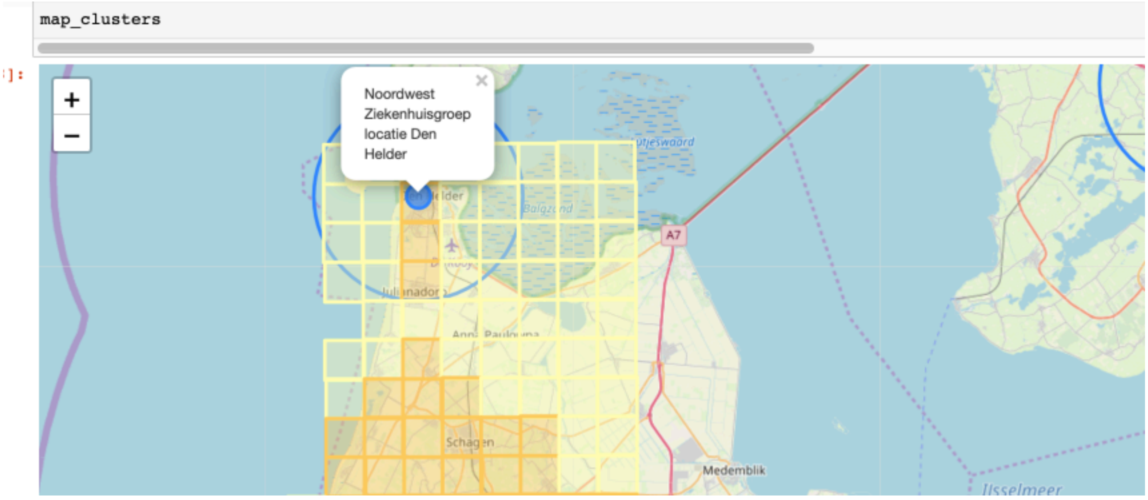


Figure 7 PM25 and Hospitals North-Holland



```
map_clusters
```

Looking at average pollution in the area called North-Holland this also shows that the northern part of this area has low pollution.
The cleanest area with also a hospital nearby is Den Helder (top left).
This is actually the area where I live now based on preliminary data, I used many years ago from different sources.

6 Discussion

Foursquare was not able to give me confidence in returning correct Hospital locations.
If you would want to make this program into a real product you would need to get a list of validated locations of hospitals.
Using foursquare for other data like Restaurants etc is ok.

7 Conclusion

For me it is clear based on the map I now produced including the hospitals I made the right choice many years ago.

I can use this program as a basis to enhance and see how I can use data like Schools, House prices etcetera.
Because of limits in the number of calls to nominatim and having a lot of data points which I needed to reduce, I might miss some low or high pollution spots.
But using 3x3 km instead of 1x1 km gives a pretty good indication of where the pollution is.