Utrecht - Air Quality Analysis

(Acquisition & Exploration of Geo Data –

Individual Assignment)

Submitted by

Pitchaporn Likitpanjamanon

(Student Number: 3234223)

Table of Contents

List of Figures

List of Tables

Introduction

Air pollution is one of the major problems that last longs for a decade. It can cause severe health issues, especially the respiratory system. To comply with WHO air quality guidelines, monitoring air pollutants such as PM 2.5 and NO_2 is mandatory. In Utrecht, Sniffer Bike project has been introduced to measure particulate matter while cycling so that cyclists can choose healthier routes (European Union, 2019). It shows how air pollution matters to people's activities like cycling which leads to the three main objectives of air quality analysis in Utrecht to help citizens gain more insight from raw data of air particle measurement. Meanwhile, the government can use these data for action plans or policy enforcement to make air quality comply with WHO guidelines. The three main objectives of this analysis are:

- How different is the air quality along bike paths during peak hours and off-peak hours?
- What are the neighborhoods of the city with the best and worst air quality along the year?
- Which areas of the city have the most and least reliable air pollution data?

Scope

In this analysis, I selected 'Utrecht' municipality as my study area and focused on PM 2.5 and NO_2 air particle measurements. The air particle measurements data were obtained from 01/01/2023 till 31/12/2023.

Folder Structure

In zip file, it contains 5 folders as follows.

- chart: pictures of plots and graphs used in the report
- data: input data used for analysis
- interpolation: interpolation result in raster format
- map: all interactive maps for visualization
- utils: Python script file contains all custom functions used in the analysis

The main code file (aq_utrecht_project.ipynb) isn't in any folder and it's in Jupyter Notebook format.

Dataset

These are the data that are used in this project.

Data	Obtained by	Owner	Description	Note
Utrecht boundary	API	PDOK	Utrecht boundary in JSON format	-
Utrecht subarea boundary	API	PDOK	Utrecht subarea boundary in JSON format	-
Bike path in Utrecht	Manually download	Geofabrik	Bike path in Utrecht province in shapefile format (data from OpenStreetMap)	Need to clip only for Utrecht municipality
Land use in Utrecht	API	PDOK	Existing land use in Utrecht in JSON format	Need to clip only for Utrecht municipality
Things (stations) data in Utrecht	API	Samenmeten	Things metadata in form of JSON format	-
Official stations data in Utrecht	Manually download	Luchtmeetnet	Official stations metadata from CSV file	Need to select stations in Utrecht municipality
PM 2.5 and NO_2 measurements	API	Samenmeten	PM 2.5 and NO_2 measurements data in JSON format	-
	Manually download	Luchtmeetnet	PM 2.5 and NO_2 measurements data in CSV file (each CSV file contains data for each month)	-

Table 1 Dataset descriptions

For data from Luchtmeetnet, I didn't fetch data from API because the endpoint that I could pass the start date and end date only returned the last 7-day records and they also suggested manually downloading data if large historical data needs to be obtained.

Steps

1. Get Data

1.1. Utrecht boundary

The response was in JSON format including municipality metadata, bounding box, and municipality boundary. The coordinate reference system (CRS) was Amersfoort / RD New (EPSG:28992). After getting the response, save the response in GeoJSON file.

1.2. Utrecht subarea boundary

The response was in JSON format including each subarea metadata and subarea boundary. The coordinate reference system (CRS) was Amersfoort / RD New (EPSG:28992). After getting the response, save the response in GeoJSON file.

1.3. Bike path in Utrecht

Bike path data in Utrecht was downloaded from Geofabrik website and the data was exported from OpenStreetMap. However, they only provided the data at the province level. So I need to clip it by using municipality boundary. Before clipping it, I already set the coordinate reference system (CRS) of the bike path shape file to Amersfoort / RD New (EPSG:28992).

1.4. Land use in Utrecht

The response was also in JSON format with Amersfoort / RD New (EPSG:28992) coordinate reference system. The land use needed to be clipped as well.

1.5. Things (stations) data in Utrecht

To filter the things that are in Utrecht, I used 'Municipality code' that I got from getting Utrecht boundary as a value of '$filter' query parameter since filtering by using spatial operation didn't work. The response was in JSON format. It contained a list of things' metadata including their name, location (coordinates), and available data streams of things.

1.6. Official stations data in Utrecht

Official stations metadata in Utrecht could obtained from PM 2.5 and NO_2 measurements CSV file. It also contained name, location, and station type.

1.7. PM 2.5 and NO_2 measurements

The PM 2.5 and NO_2 measurements data from both Samenmeten and Luchtmeetnet were reported hourly. I took the data of year 2023 to do the analysis (01/01/2023 – 31/12/2023).

1.7.1. **Data from Samenmeten**: I obtained the measurements of each things by using data stream observation's link that I got from getting Things data. Data were processed and formatted in form of tabular and I saved them as csv file.

Note that there were PM 2.5 raw data and calibrated PM 2.5 data. I selected **calibrated PM 2.5 data** for doing this analysis because it was more accurate than the raw data.

	phenomenonTime	AMF_pm012	GLB_1745609	GLB_4182551	GLB_5939801	GLB_5965467	GLOBE_350916063264611
0	2023-01-01T00:00:00.000Z	30.932	NaN	NaN	58.52	NaN	NaN
1	2023-01-01T04:00:00.000Z	4.112	NaN	NaN	3.00	NaN	NaN
2	2023-01-01T05:00:00.000Z	2.100	NaN	NaN	3.05	NaN	NaN
3	2023-01-01T06:00:00.000Z	4.425	NaN	NaN	2.16	NaN	NaN
4	2023-01-01T09:00:00.000Z	1.980	NaN	NaN	2.00	NaN	NaN
...	...	...	...	...	...	...	...
7701	2023-12-31T19:00:00.000Z	5.876	NaN	NaN	NaN	2.49035	3.85172
7702	2023-12-31T20:00:00.000Z	7.263	NaN	NaN	NaN	17.77420	11.97250
7703	2023-12-31T21:00:00.000Z	8.730	NaN	NaN	NaN	5.28801	6.32022
7704	2023-12-31T22:00:00.000Z	8.811	NaN	NaN	NaN	18.46090	10.16980
7705	2023-12-31T23:00:00.000Z	10.146	NaN	NaN	NaN	2.63084	4.06342

Figure 1 Example of data structure (Samenmeten)

Data structure of both PM 2.5 and NO_2 measurements consisted of the following columns:

- phenomenonTime: Time that the observation is taken (hourly). It is in UTC time zone.
- The rest of the columns are Things' names which contain observation values of that Thing.

1.7.2. **Data from Luchtmeetnet**: One csv file contained the data for one month so I need to aggregate all data for 12 files and store it in dataframe.

	Component	Bep.periode	Eenheid	Begindatumtijd	Einddatumtijd	NL10639	NL10636	NL10643
0	NO2	uur	µg/m³	20230101 00:00	20230101 01:00	24.05	15.42	14.95
1	NO2	uur	µg/m³	20230101 01:00	20230101 02:00	10.75	9.31	7.95
2	NO2	uur	µg/m³	20230101 02:00	20230101 03:00	6.00	5.46	4.97
3	NO2	uur	µg/m³	20230101 03:00	20230101 04:00	4.63	4.75	4.18
4	NO2	uur	µg/m³	20230101 04:00	20230101 05:00	4.39	3.35	3.40
...	...	...	...	...	...	...	...	...
8755	NO2	uur	µg/m³	20231231 19:00	20231231 20:00	9.62	7.68	6.97
8756	NO2	uur	µg/m³	20231231 20:00	20231231 21:00	8.42	8.16	7.11
8757	NO2	uur	µg/m³	20231231 21:00	20231231 22:00	7.24	6.91	5.86
8758	NO2	uur	µg/m³	20231231 22:00	20231231 23:00	7.92	7.36	6.23
8759	NO2	uur	µg/m³	20231231 23:00	20240101 00:00	7.48	6.29	5.69

Figure 2 Example of data structure (Luchtmeetnet)

Data structure of both PM 2.5 and NO_2 measurements consisted of the following columns:

- Component: the air particle
- Bep.periode: Time interval (hourly)
- Eenheid: measurement unit
- Begindatumtijd: start date (UTC+1 time zone)
- Einddatumtijd: end date (UTC+1 time zone)
- The rest of the columns are stations' names which contain observation values of that station.

In this analysis, I picked 'Begindatumtijd' (start date) column as a reference for measurement value because it made more sense. If I selected Einddatumtijd column, I wouldn't have data at 2023-01-01 (00:00) and have 2024-01-01 (00:00) instead.

2. Time zone conversion

I did the time zone conversion first to make sure that data from both sources were in the same time zone so the analysis result became accurate.

2.1. Samenmeten

By default, the time column (phenomenonTime) was in the UTC time zone so I converted it to 'Europe/Amsterdam' time zone. It automatically handled the daylight saving time.

2.2. Luchtmeetnet

The time columns (Begindatumtijd and Einddatumtijd) were in the UTC+1 time zone so I need to localize these columns to UTC+1 time zone. Then, I converted them to 'Europe/Amsterdam' time zone.

3. Make sure that data from Samenmeten and Luchtmeetnet have the same length (number of rows)

For one year data, measurements data for each station should have 8760 records. Data from both sources had some missing values (NaN value). I found that data from Luchtmeetnet already had 8760 records, unlike Samenmeten. I needed to insert dates and times that were missed from the obtained data and put the value as NaN. It would be easier to compare or merge data when the data had the same length.

4. Explore Data

4.1. Samenmeten

I found that there were inconsistencies between the location of some 'Things' in Samenmeten data portal and data from API. Hence, I decided to inspect the network tab from the browser to find the API that the portal used. Then, I compare the location coordinates from both sources and update new coordinates if they were mismatched.

API: https://samenmeten.rivm.nl/dataportaal/php/getData-fromfile.php?compartiment=lucht

Then, I explored the owner of each thing and the sensor type used for measuring PM 2.5 and NO_2. The information of each owner and sensor type is listed below.

Owner information

Owner category	Owner	Description	Note
Government	RIVM	National Institute for Public Health and the Environment	-
	Apeldoorn	Apeldoorn Municipality	-
Professional Organization	Flexyz	IT company and IoT is one of their services as well	-

	Globe	Citizen Science Project that partner with RIVM to focus on school education	I put it in professional organization category because they partner with RIVM
	UU	University of Utrecht	-
	Smart Emission	Research Project partners with governments and citizens in Nijmegen	-
Citizen Science Project	Hollandse Luchten	Citizen Science Project focused on the environment measurement in North Holland	-
	Luftdaten (Sensor Community)	Citizen Science Project which is on an international scale	-
	Meetjestad	Citizen Science Project	-
Others	NB-IoT	Narrowband IoT	I wasn't sure about these three owners so I put them as 'Others' category
	Maarten vd B	This might be some person's name from searching in Google	
	Palmes	Probably it is palmes tube for NO_2 measurement	

Table 2 Owner category

Sensor information

PM 2.5

Quality Rank	Sensor
High	Sensirion SPS030
Medium	Plantower PMS7003
Low	Nova SDS011

Table 3 PM 2.5 Sensor quality rank

According to Samenmeten website, they indicated that SDS011 is more suitable for detecting coarser particles compared to PMS7003 and SPS030 (RIVM, n.d.). Therefore, I put it as the lowest rank since this study focused on PM 2.5, not PM 10. For SPS030 and PMS7003, I checked their datasheet and found that SPS030 had a bit better higher mass concentration precision for PM 2.5 and longer mean-time-to-failure (MTTF) as shown in Table 4.

Property	Sensirion SPS030	Plantower PMS7003
Mass concentration precision for PM 2.5	0 to 100 µg/m^3 ±10 µg/m3	0 to 100 µg/m^3 ±10 µg/m3
	100 *to 1000* µg/m^3 ±10% m.v.	100 *to 500* µg/m^3 ±10% m.v.
MTTF	>= 3 years	> 10 years

Table 4 SPS030 and PMS7003 Comparison

NO_2

Quality Rank	Sensor
-	Alphasense NO_2

Table 5 NO_2 sensor quality rank

For NO_2, there was only one sensor which is Alphasense NO_2.

Data Exploration (focused on Things information)

In this exploration, I used spatial join (intersect) to join existing land use, Things, and Utrecht subarea boundary geodataframes together.

Things related to PM 2.5 measurement

There were 95 Things used for PM 2.5 measurement in 2023.

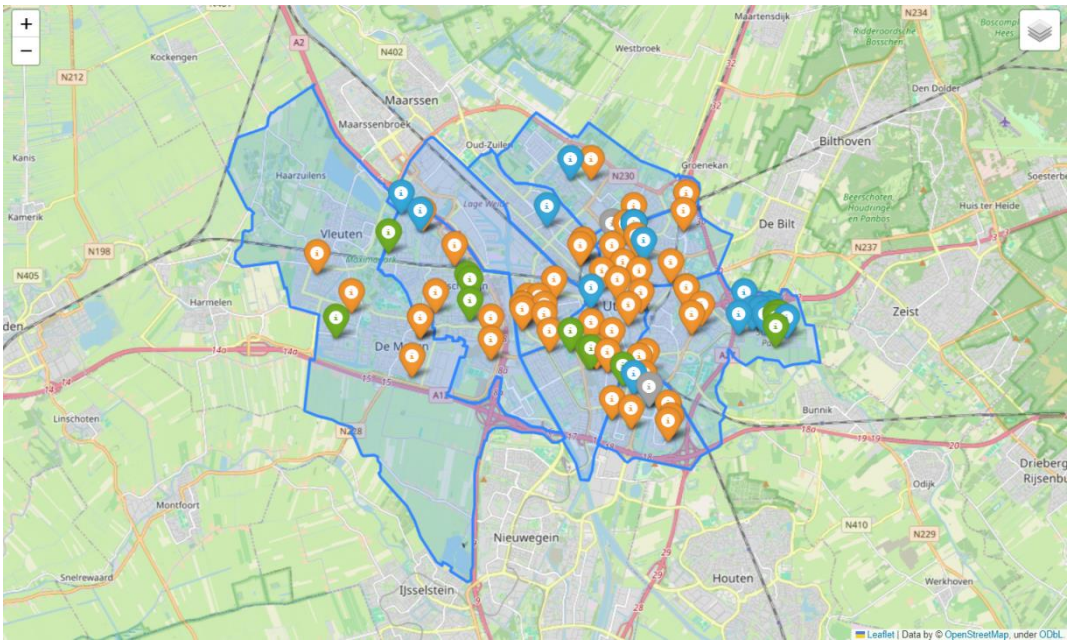

Figure 3 Spatial Distribution of Things (PM 2.5 measurement)

From Figure 3, all Things were placed all over Utrecht. The color shows the owner category of each Thing based on Table 6.

Color	Owner category
Blue	Government
Green	Professional Organization
Orange	Citizen Science Project
Gray	Others

Table 6 Color of each owner category

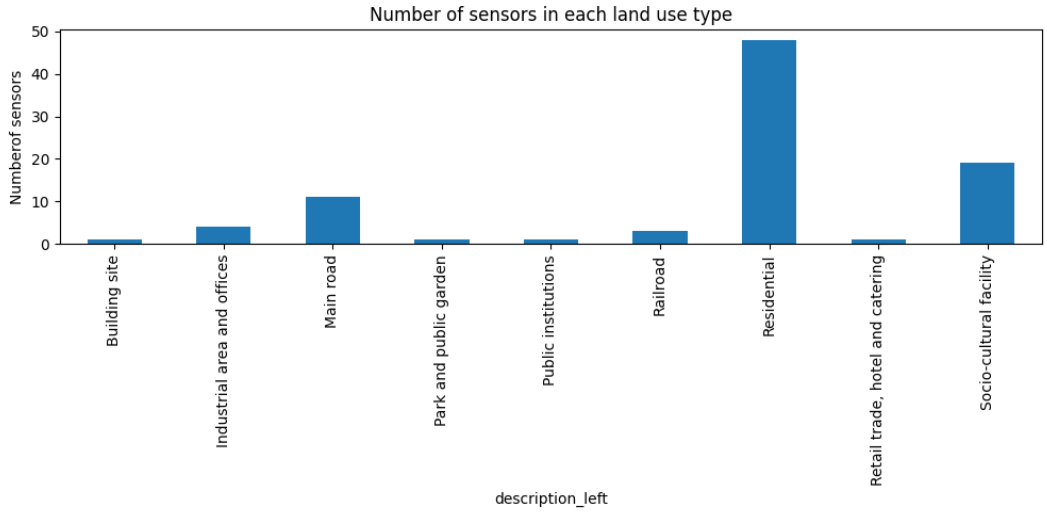

Figure 4 Number of sensors in each land use type

Most of the PM 2.5 sensors were placed in residential areas and only one sensor was placed in building site, park and public garden, public institutions, and retail trade, hotel, and catering area.

Things related to NO_2 measurement

There were 19 Things used for NO_2 measurement in 2023.

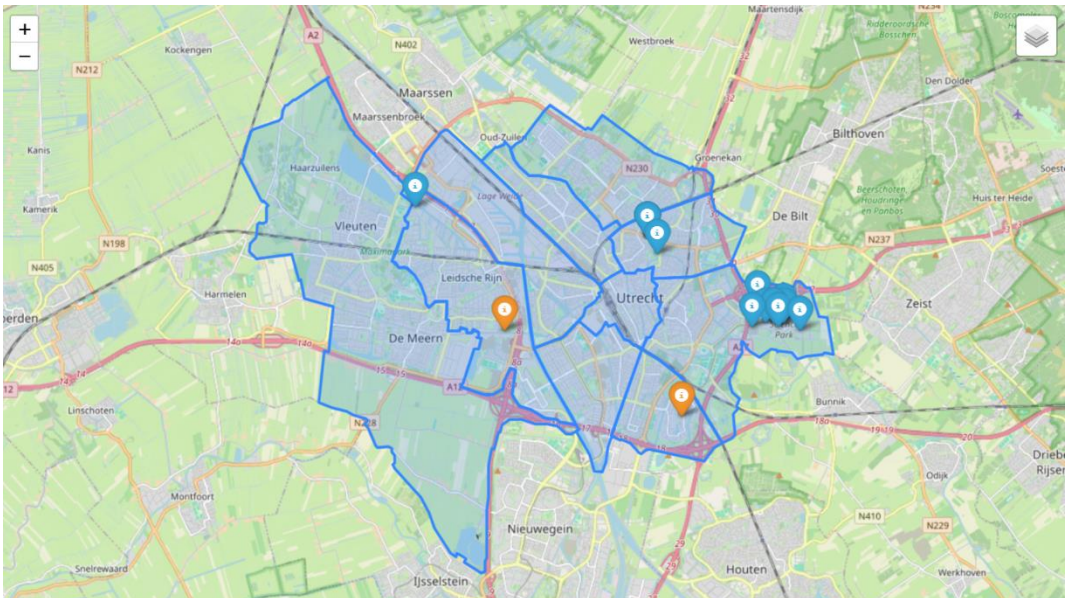

Figure 5 Spatial Distribution of Things (NO2 measurement)

Unlike PM 2.5, Things were mostly placed in the eastern of Utrecht not all over Utrecht as shown in Figure 5. Please refer to Table 6 for the Marker's color meaning. The owners are the Government and Citizen Science Project.

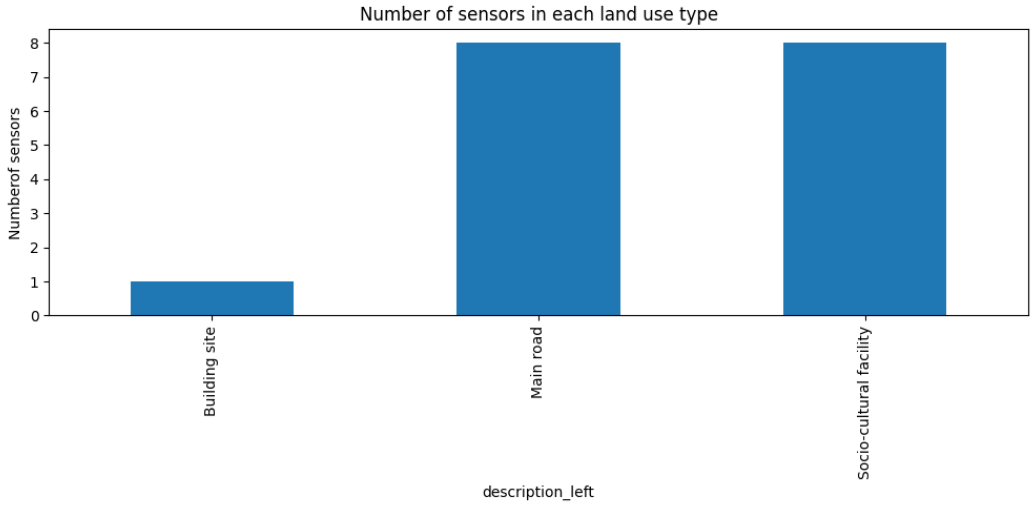

Figure 6 Number of sensors in each land use type

Most NO_2 sensors were placed on or near the main road and socio-cultural facility, but none of them were placed in residential areas.

More findings were discussed in the *Result* section under *Which areas of the city have the most and least reliable air pollution data?* Explanation.

4.2. Luchtmeetnet

Data Exploration (focused on Station information)

I got the official station codes in Utrecht by getting unique values from *no2closecode* and *pm25closecode* attributes in Things metadata. Then, I filtered official stations data by using list of station codes.

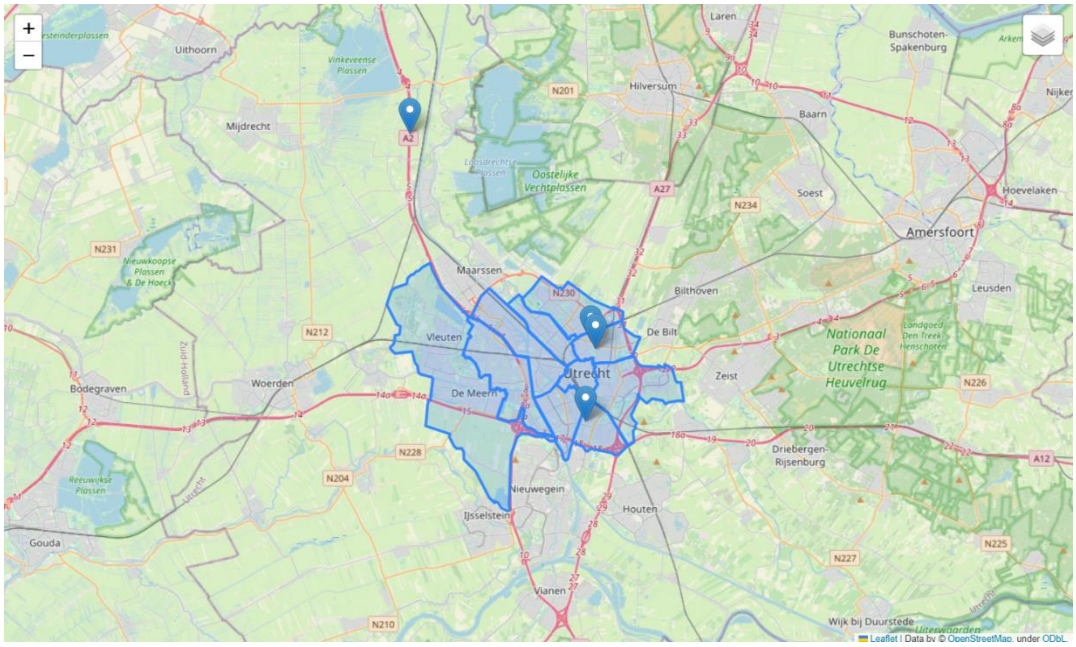

Figure 7 Spatial Distribution of official stations

All of official stations owned by RIVM. However, there was no information about sensors used for air particle measurements. From Figure 7, there was one station that was out of Utrecht so I filtered it out. In total, data from three stations were used for analysis. Two stations were street station and another one was a city background station. For PM 2.5 measurements, data were from NL10636 and NL10643 station, and for NO_2, data were obtained from all three stations.

5. Explore Data Distribution and relationship between PM 2.5 and NO_2

From this section onwards, I used 'station' instead of 'Thing' so it could be easier to understand.

5.1. Samenmeten Data

For both PM 2.5 and NO_2 data, I applied *.describe()* function to get descriptive statistics (data points, mean, std, min, max, Q1, Q2, and Q3) of data and started to inspect more on each dataset. To inspect both datasets, I selected PM 2.5 and NO_2 measurements from HLL_zps-09 station to visualize the histogram, boxplot, and relationship between PM 2.5 and NO_2.

PM 2.5 Data

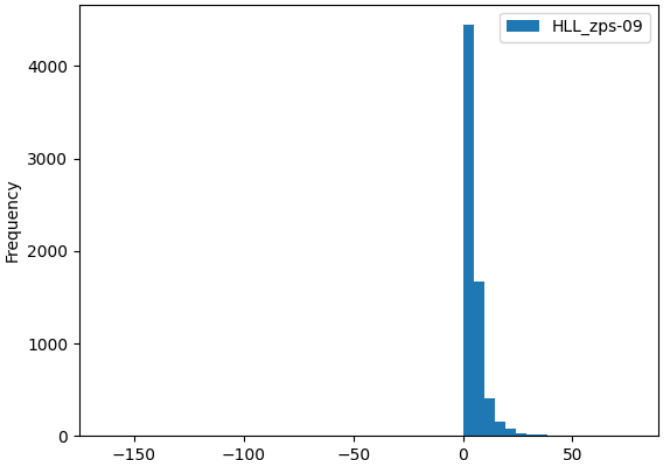

Figure 8 Histogram of PM 2.5 data from HLL_zps-09 station

From the histogram, data from HLL_zps-09 station has right skewed and it had the outliers because the x-axis has a minus value. I plotted the boxplot to verify this assumption and see outliers. Boxplot revealed that this data had outliers.

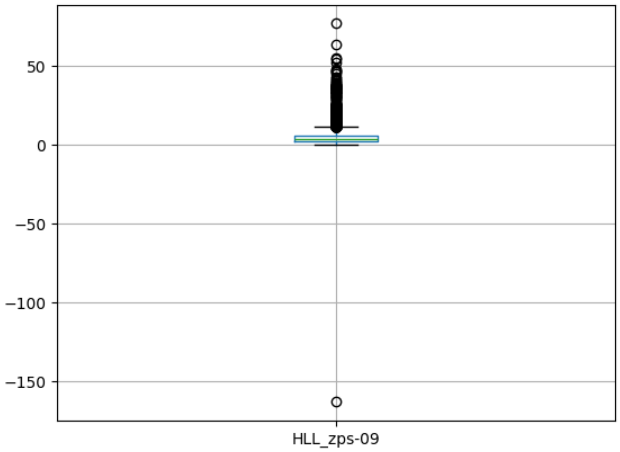

Figure 9 Boxplot of PM 2.5 data from HLL_zps-09 station

Then, I applied *.skew()* function to check the skewness of data from other stations. Most of them (data from 88 stations) had **right-skewed**, but there was one station (HLL_zps-10) that had normal distribution. I found that data from that station had only 0 as its value. For outliers, I used *IQR method* (formula were shown below) to detect outliers in data. The result was data from other stations also had outliers as well.

$$IQR = Q3 - Q1$$
$$lower\ bound = Q1 - 1.5 * IQR$$
$$upper\ bound = Q3 + 1.5 * IQR$$

NO_2 Data

By applying *.describe()*, I found -999 value which indicated no data so I replaced -999 with NaN first before checking data distribution.

The same steps also applied to NO_2 data as well.

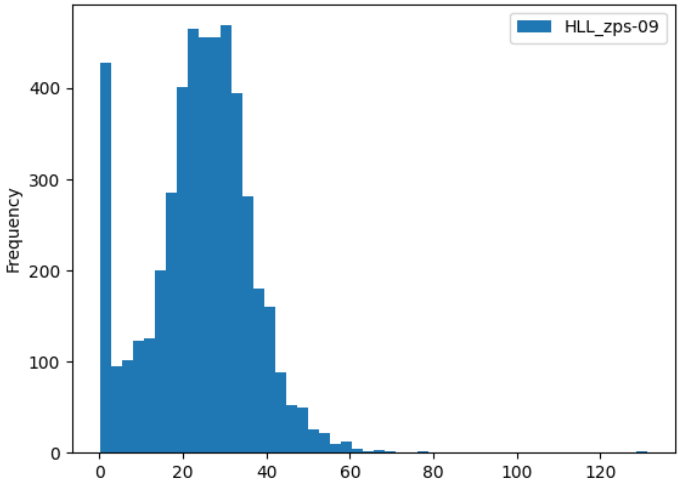

Figure 10 Histogram of NO_2 data from HLL_zps-09 station

The histogram showed that NO_2 data also had right skewed as well as PM 2.5 data and boxplot also showed the outliers in data.

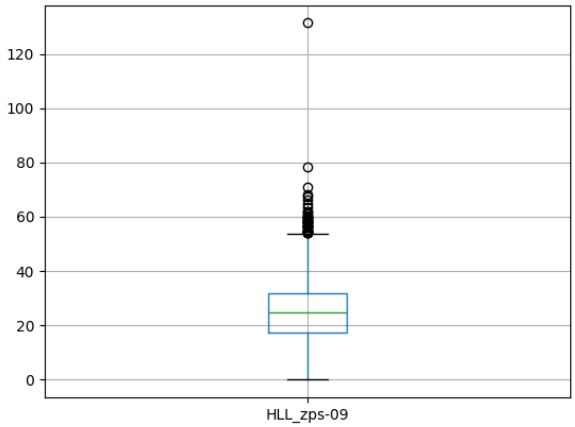

Figure 11 Boxplot of NO_2 data from HLL_zps-09 station

After checking the skewness of data from other stations, all of them had **right-skewed**. For outliers detection, *IQR method* was also applied and most of them had outliers as well. I also found that there were stations that contained only NaN value (after replacing -999 value with NaN) which needed to be removed.

Relationship between PM 2.5 and NO_2

Additionally, I wanted to explore the relationship between PM 2.5 and NO_2 whether changes in one air particle variable would affect another air particle variable or not. I used scatter plot and linear regression to observe their relationship or correlation. To evaluate their relationship, I used *r value* which represented the correlation between them.

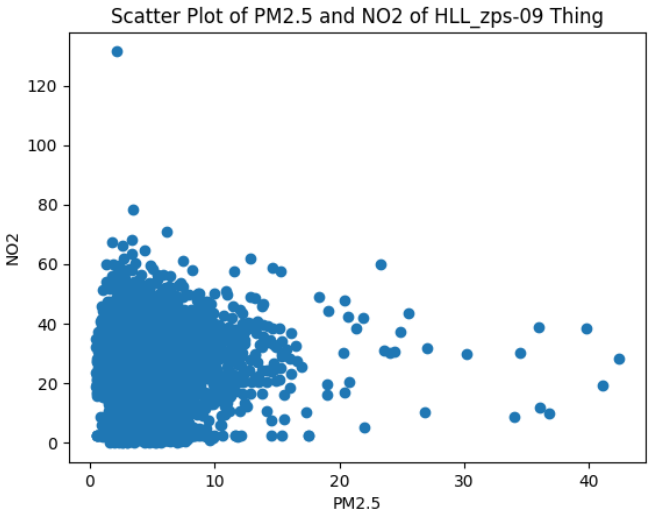

Figure 12 Scatter plot between PM 2.5 and NO_2 (Samenmeten)

From the scatter plot, I couldn't see a linear relationship between PM 2.5 and NO_2, but it needed to be confirmed by r value again so I applied linear regression by using *scipy.stats.linregress()* function. The **r value that I got was 0.021** which indicated that **PM 2.5 and NO_2 had no relationship**. (Note: This is just the data from one station)

5.2. Luchtmeetnet Data

I repeated the same steps that I did with Samenmeten data. For Luchtmeetnet data, I selected station NL10636 which had both PM 2.5 and NO_2 measurements to visualize histogram, boxplot, and relationship between PM 2.5 and NO_2 data. For outliers detection, IQR method also applied to Luchtmeetnet data as well.

PM 2.5 Data

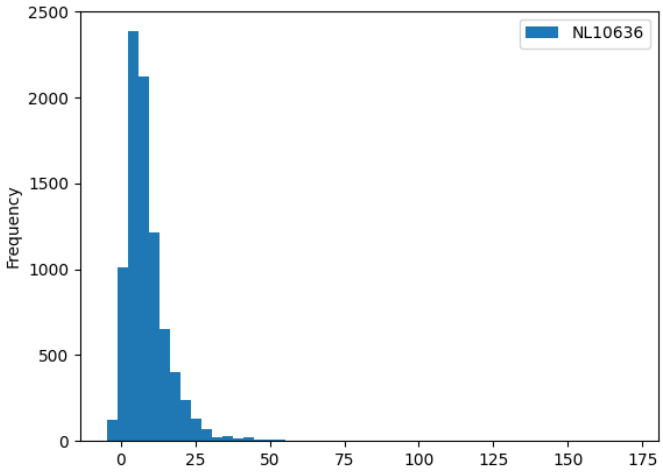

Figure 13 Histogram of PM 2.5 data from NL10636 station

PM 2.5 data from NL10636 station also had right-skewed like PM 2.5 data from Samenmeten and it also had outliers as shown in Figure 11.

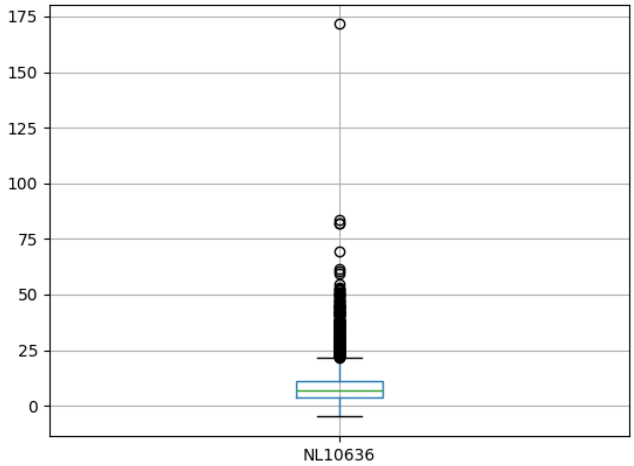

Figure 14 Boxplot of PM 2.5 data from NL10636 station

I found that data from all stations had **right-skewed** and outliers.

NO_2 Data

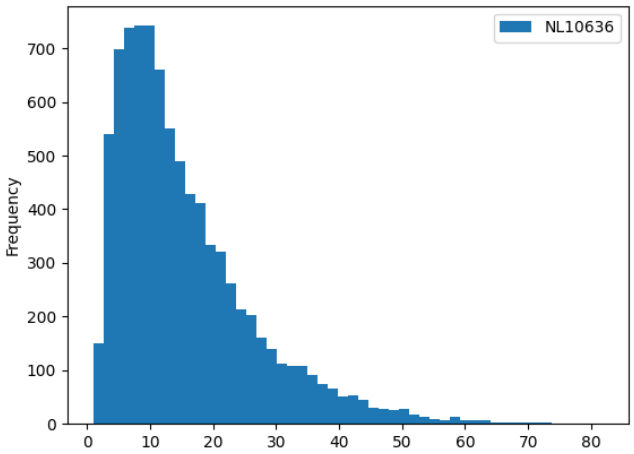

Figure 15 Histogram of NO_2 data from NL10636 station

From above histogram, NO_2 data also had right-skewed as well. The outliers were clearly shown in boxplot.

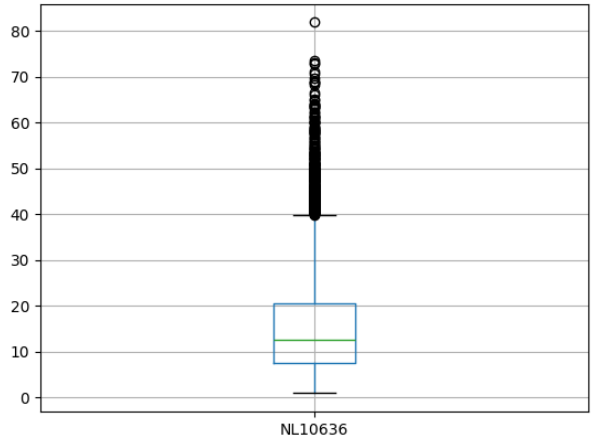

Figure 16 Boxplot of NO_2 data from NL10636 station

NO_2 data from all three stations had **right-skewed** and outliers. I could say that both PM 2.5 and NO_2 data from Luchtmeetnet and Samenmeten had the right-skewed data distribution and both had outliers in data.

Relationship between PM 2.5 and NO_2

I also applied scatter plot and linear regression to find the relationship between PM 2.5 and NO_2 in Luchtmeetnet dataset.

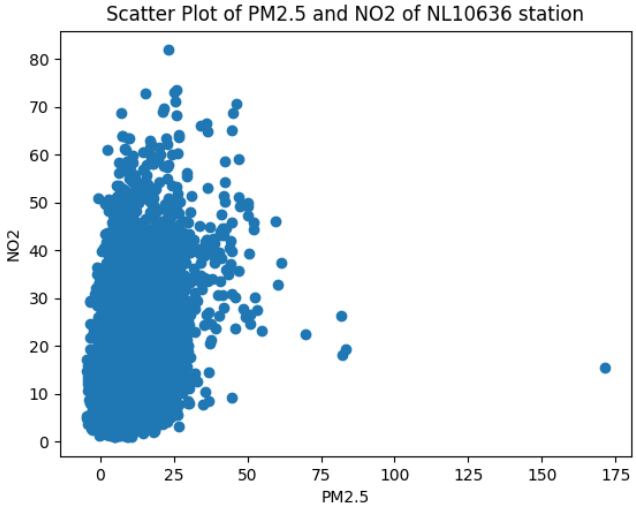

Figure 17 Scatter plot between PM 2.5 and NO_2 (Luchtmeetnet)

From the scatter plot, it seemed to have more pattern in their relationship compared to Samenmeten data. After applying linear regression, the **r value (correlation coefficient) was 0.379**. PM 2.5 and NO_2 measurements from Luchtmeetnet had a positive linear relationship, which means when PM 2.5 increases, NO_2 tends to increase as well. However, 0.379 is still considered as a weak relationship so a positive relationship isn't clearly shown. In contrast, Samenmeten data showed no relationship between PM 2.5 and NO_2 and this might happen because of the data quality. Data from Luchmeetnet were cleaner and had fewer missing values than Samenmeten data.

6. Clean Data

6.1. Samenmeten Data

PM 2.5 Data

- Removed HLL_zps-10 column since the value from this station was only 0
- Replaced minus value with NaN value
- Replaced the value that exceeded the maximum reading of the sensors with NaN

  I found that data from SPS030 and SDS011 exceeded their measurement range (0-1000 $\mu g/m^3$).
- Clean duplicated stations (The station had the same location)

  I needed to clean them because I used the interpolation in the data analysis step each location should have one set of data. I cleaned them by counting the number of data points of each duplicated station and kept the station that had the highest number of data points from each location. Here is an example.

Station	Location	Data points
A	Location 1	1000
B	Location 1	4000
C	Location 2	1500
D	Location 2	4500

Table 7 Example of duplicated station data

In this case, data from Station B and Station D are kept and data from Station A and Station B are removed.

- Remove stations (columns) that had fewer data points

  From descriptive statistics, I found that some stations (columns) had only one or few data points so I created two criteria for cleaning them. One was for data that was used for the annual average analysis and another one for peak and off-peak hour analysis.

  - Annual average analysis: removed station that had number of data points less than 5% of ideal data points (8760 records)
  - Peak and off-peak hour analysis: removed station that had number of data points less than 100 data points

  I split it into two criteria because to do the annual average calculation, it is better to have more number of data points as it should represent the whole year information. Meanwhile, off-peak and peak hour average are in smaller scale so I reduced the threshold value for filtration.

NO_2 Data

- Clean duplicated stations (The station had the same location)

  The same logic as described previously was applied to NO_2 data as well.

- Remove stations (columns) that had less data points

  The same logic as described previously was applied to NO_2 data as well.

6.2. Luchtmeetnet Data

There were minus values in PM 2.5 measurements as well. However, according to the official website page, they mentioned that minus values in the report are acceptable since they are in an acceptable range so they should be kept for analysis.

7. Data Analysis

7.1. How different is the air quality along bike paths during peak hours and off-peak hours?

Peak and Off-peak hours

I categorized peak and off-peak hours based on the information from the NS website.

Peak hours: On weekdays from 6.30 a.m. to 9 a.m. and 4 p.m. to 6.30 p.m. Other than that is off-peak. Weekend and Holiday are also counted as off-peak hours (Weekend is Friday evening 6.30 p.m. to Monday morning 4.00 a.m.).

However, the time reported in data was hourly so it wasn't possible to check for 6.30 a.m./p.m. So I used 6 am. and 7 pm. instead.

After defined peak and off-peak hours, I applied the steps below (All steps were applied to both PM 2.5 and NO_2 data).

Geodataframe Preparation

- Joined two dataframes (data from Samenmeten and Luchtmeetnet) by using datetime column
- Filtered data within peak and off-peak hours and created two new dataframes to keep data separately (peak and off-peak dataframes). Besides from day in a week and hour, I also defined a list of The Netherlands' public holidays in 2023 to use in filtering as well.
- Calculate **average of PM 2.5/NO_2** for each station in peak and off-peak dataframes
- Created geodataframes corresponding to peak and off-peak dataframes and created point geometry (latitude and longitude coordinate in dataframes were in EPSG:4326)
- Converted CRS of geodataframes to Amersfoort / RD New (EPSG:28992) since it is better to use the projected coordinates system for distance calculation
- Extract projected coordinates (x, y) from point geometry and store in 'Easting' and 'Northing' column
- Got peak and off-peak geodataframes which were used as input for interpolation

Interpolation

PM 2.5

There are many interpolation method, but I chose 'Linear Interpolation' method for interpolating PM 2.5 and NO_2 value because it's suitable for continuous data while nearest neighbor is more suitable for discrete data. Furthermore, there were linear relationships between PM 2.5/ NO_2 value from each station and this can be found by using correlation coefficient value.

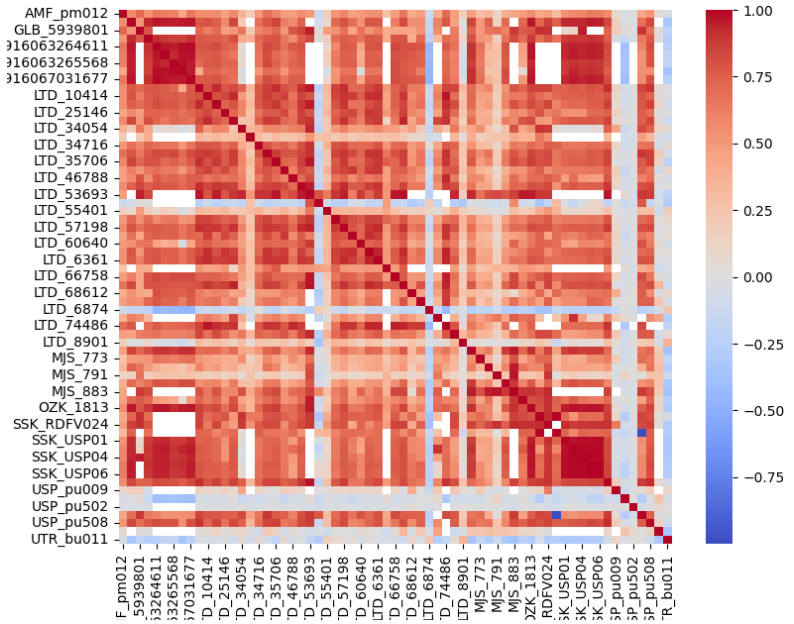

Figure 18 Correlation Heatmap (PM 2.5 stations from Samenmeten)

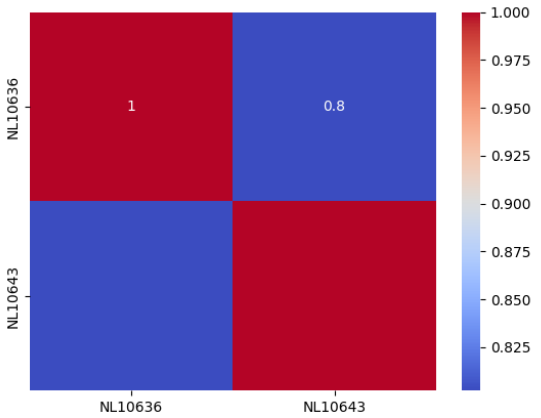

Figure 19 Correlation Heatmap (PM 2.5 stations from Luchtmeetnet)

From the correlation heatmap in Figure 18 and 19, the linear relationship could be seen and most of them were moderate and strong relationship. For the threshold, correlation coefficient > 0.5 is considered as moderate to strong relationship (Jaadi, 2019). I also checked the correlation between PM 2.5 data from stations in Samenmeten and Luchtmeetnet. 78.46% of them had moderate to strong relationship.

NO_2

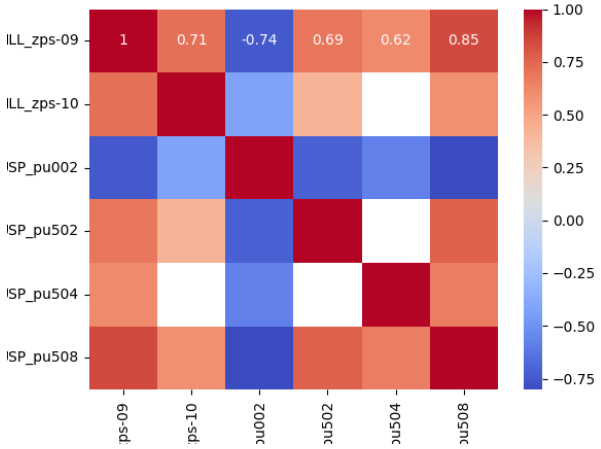

Figure 20 Correlation Heatmap (NO_2 stations from Samenmeten)

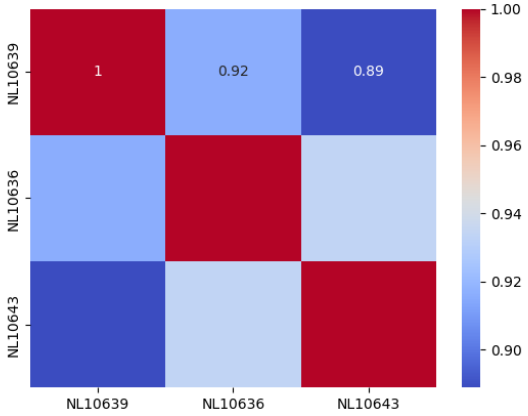

Figure 21 Correlation Heatmap (NO_2 stations from Luchtmeetnet)

For NO_2 data for both Samenmeten and Luchtmeetnet, data from different stations had a moderate to strong relationship and 50% of the correlation between NO_2 data from Samenmeten stations and Luchtmeetnet stations was moderate to strong relationships.

Finding the difference of PM 2.5/ NO_2 value along the bike path during peak and off-peak hours

After getting the two interpolation results in raster format (peak and off-peak hours), I clipped the result by using municipality boundary and rasterizing the bike path vector by assigning burn value to be 1. Then, I could assign PM 2.5/ NO_2 value to bike path by multiply the interpolation result to bike path (bike path's pixel value is 1).

Now I had PM2.5/ NO_2 value along the bike path for peak and off-peak hours in form of NumPy array so I can find the difference between them.

$$difference\ between\ off\ peak\ and\ peak\ hours = off\ peak\ hour\ value - peak\ hour\ value$$

Visualization

For visualization, I used foilum library to visualize the result in form of an interactive map. I decided to use an interactive map because it was more convenient to investigate the data in detail. For example, if there was a small change in value that caused a slight change in the represented color, users could zoom in to inspect it. Moreover, I was able to overlay the result on the base map and this could help users identify the location or area easily. The results were discussed in 'Result' section.

7.2. What are the neighborhoods of the city with the best and worst air quality along the year?

Geodataframe Preparation

- Calculated the **annual average of PM 2.5/ NO_2 value** of each station (both Samenmeten and Luchtmeetnet data)
- Concatenated annual average of PM 2.5/ NO_2 value from Samenmeten and Luchtmeetnet (they could be concatenated because they had same dimension) and converted the result to dataframe
- Created geodataframes corresponding to annual average of PM 2.5/ NO_2 value dataframe and created point geometry (latitude and longitude coordinate in dataframes were in EPSG:4326)
- Converted CRS of geodataframes to Amersfoort / RD New (EPSG:28992) since it is better to use the projected coordinates system for distance calculation
- Extract projected coordinates (x, y) from point geometry and store in 'Easting' and 'Northing' column
- Got annual average geodataframes which were used as input for interpolation

Interpolation

The interpolation method that was used to interpolate the annual average of PM 2.5/ NO_2 value was also the same as the interpolating average value during peak and off-peak hours.

Categorized PM 2.5 and NO_2 measurements value

According to Luchtmeetnet page, they provided the below picture which defines the threshold for each air particle measurement used for defining the air quality level.

LUCHTKWALITEIT	LKI [index]	NO_2 [µg/m³]	O_3 [µg/m³]	PM10 [µg/m³]	PM2.5 [µg/m³]
GOED	0-1	0-10	0-15	0-10	0-10
	1-2	10-20	15-30	10-20	10-15
	2-3	20-30	30-40	20-30	15-20
MATIG	3-4	30-45	40-60	30-45	20-30
	4-5	45-60	60-80	45-60	30-40
	5-6	60-75	80-100	60-75	40-50
ONVOLDOENDE	6-7	75-100	100-140	75-100	50-70
	7-8	100-125	140-180	100-125	70-90
SLECHT	8-9	125-150	180-200	125-150	90-100
	9-10	150-200	200-240	150-200	100-140
ZEER SLECHT	>10	>200	>240	>200	>140

Figure 22 Threshold value for each air particle measurement corresponding to air quality level

For this task, I assigned the air quality level by referring to Figure 22. I assigned the air quality level into two scales.

- Overview of air quality level

Air quality level	PM 2.5	NO_2	Assigned value
Good	0-20	0-30	1
Moderate	20-50	30-75	2
Poor	50-90	75-125	3
Bad	90-140	125-200	4
Very Bad	>140	>200	5

Table 8 Overview of air quality level criteria

- Identifying the area that has the best air quality regarding PM 2.5/ NO_2 value

Air quality level	PM 2.5	NO_2	Assigned value
Good (best)	0-10	0-10	1
Good (second)	10-15	10-20	2
Good (third)	15-20	20-30	3
Moderate	20-50	30-75	4
Poor	50-90	75-125	5
Bad	90-140	125-200	6
Very Bad	>140	>200	7

Table 9 Criteria for Identifying the area that has the best air quality regarding PM 2.5/ NO_2 value

Note that in my code, I assigned the float number instead of the integer due to limitations on folium visualization.

Visualization

For visualization, I also used foilum library to visualize the result in form of an interactive map. Interactive map was useful in this case since it allowed users to identify the area that has the best and worst air quality easily by overlaying the raster on the base map. The results were discussed in 'Result' section.

7.3. Which areas of the city have the most and least reliable air pollution data?

Prepare geodataframe

For stations' metadata from Samenmeten, it was already prepared during the cleaning data step. For stations' metadata from Luchtmeetnet, I used spatial joined (intersect) the original dataframe with Utrecht subarea dataframe.

Use of .groupby() function

To perform this task, I used .groupby() function to group data based on different columns so I could see data in many aspects and could be used to decide reliability of air quality data.

Visualization

I used an interactive map to visualize the spatial distribution of stations across Utrecht and also bar chart to visualize data from applying .groupby() function. Interactive map made users see spatial distribution of stations on different scale at a time by using zoom-in or zoom-out to focused area and also detail of each station by clicking on Markers in the map. For the rest, I chose the bar chart to visualize the data because it was easy to interpret and understand. It also allowed users to see the comparison of more than one variable aspect like a grouped bar chart which was shown in 'Result' section.

Result

Disclaimer: For question 1 and 2, the interpolation result didn't cover all areas in Utrecht so I interpreted based on the covered area only.

1. How different is the air quality along bike paths during peak hours and off-peak hours?

PM 2.5

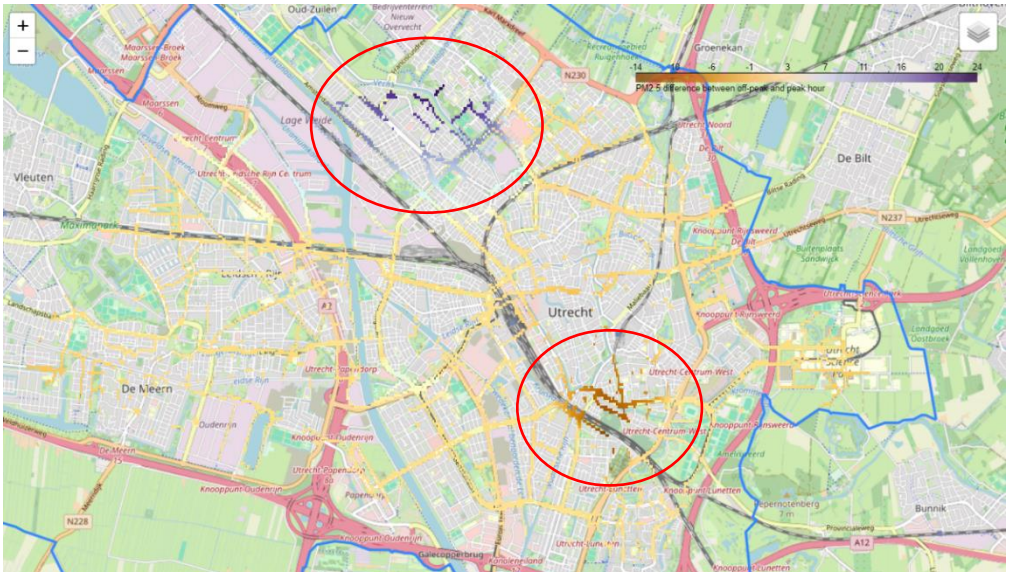

Figure 23 PM 2.5 difference between off-peak and peak hour

Url to map in Figure 23: https://gisedu.itc.utwente.nl/student/s3234223/airquality-analysis/map/pm25_difference.html

Figure 23 shows that most bike path has few differences in PM 2.5 value, but two spots had major differences. The first spot is the bike path in the north of Utrecht. For this spot, PM 2.5 during off-peak hours is higher than during peak hours (maximum difference is approximately 24 µg/m^3). The second area is down to the south of Utrecht. Unlike the first area, PM 2.5 along the bike path during peak hours is higher than during off-peak hours (maximum difference is approximately 14 µg/m^3) which makes sense since it is near the centrum.

Url to average PM 2.5 during peak and off-peak hours

Peak hour: https://gisedu.itc.utwente.nl/student/s3234223/airquality-analysis/map/pm25_peak_hour.html

Off-peak hour: https://gisedu.itc.utwente.nl/student/s3234223/airquality-analysis/map/pm25_offpeak_hour.html

NO_2

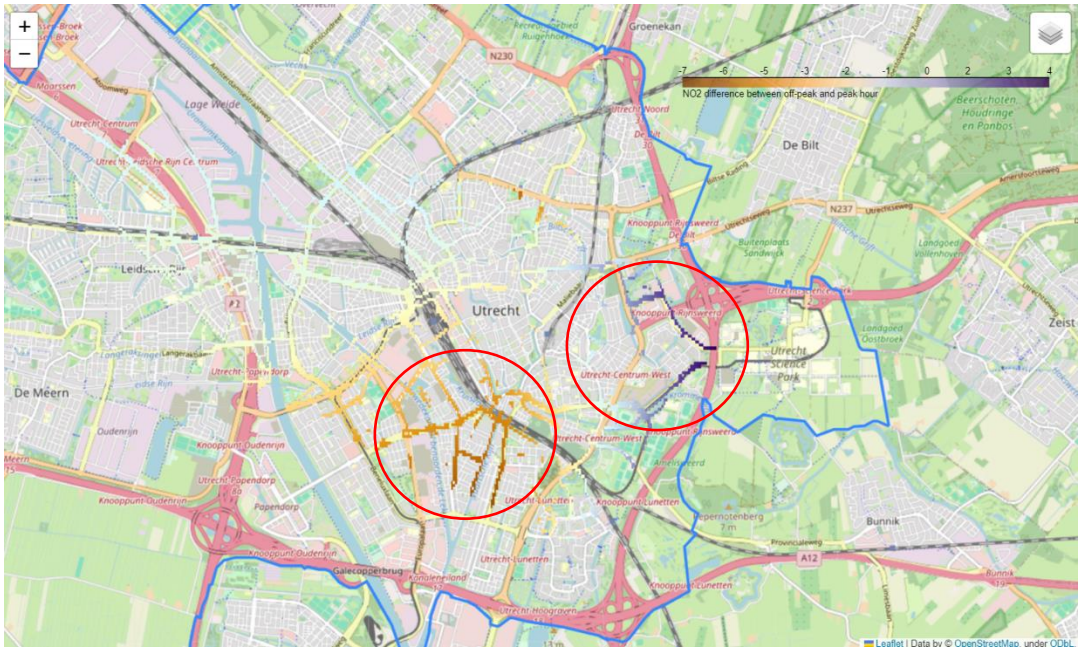

Figure 24 NO_2 difference between off-peak and peak hour

Url to map in Figure 24: https://gisedu.itc.utwente.nl/student/s3234223/airquality-analysis/map/no2_difference.html

For NO_2 value, the covered bike path area is less than PM 2.5 result because the stations for NO_2 measurement are less than PM 2.5. Most covered areas are on the east side and down to the south of Utrecht. Two areas show the difference between NO_2 values during off-peak and peak hours. Both are also near the centrum, but one area has a higher NO_2 value during off-peak hours than during peak hours (maximum difference is around 4 µg/m^3). In contrast, NO_2 is higher during peak hours than during off-peak hours in another area (maximum difference is around 7 µg/m^3). This area is near the Utrecht Central Station and from the map, bike path in this area is near the railway station as well.

Url to average NO_2 during peak and off-peak hours

Peak hour: https://gisedu.itc.utwente.nl/student/s3234223/airquality-analysis/map/no2_peak_hour.html

Off-peak hour: https://gisedu.itc.utwente.nl/student/s3234223/airquality-analysis/map/no2_offpeak_hour.html

2. **What are the neighborhoods of the city with the best and worst air quality along the year?**

PM 2.5

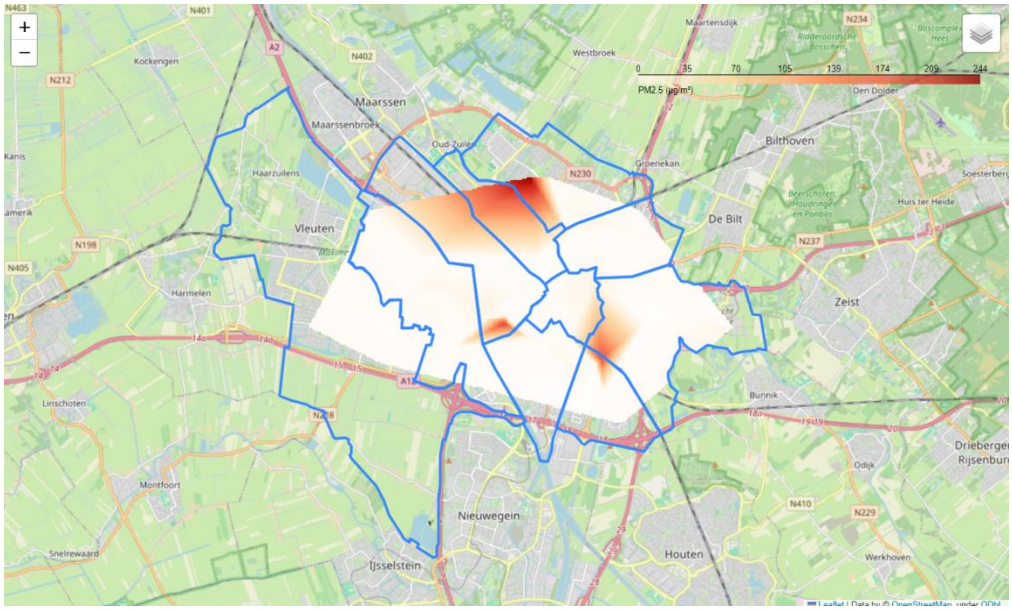

Figure 25 Annual average of PM 2.5 value

Url to map in Figure 25: https://gisedu.itc.utwente.nl/student/s3234223/airquality-analysis/map/pm25_annual_average.html

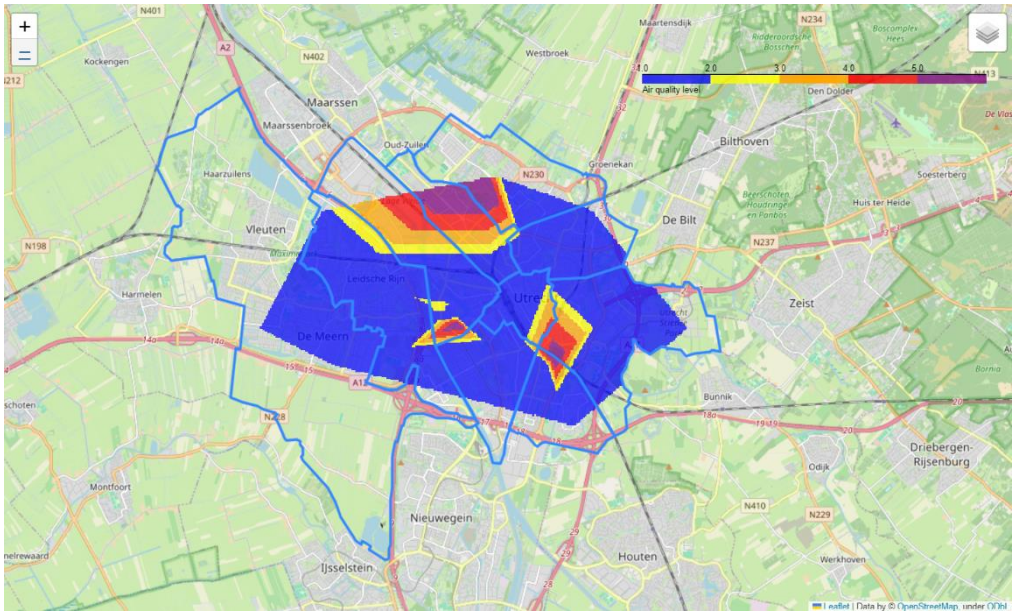

Figure 26 Air quality level based on criteria in Table 8

Figure 25 shows the annual average PM 2.5 value from the linear interpolation method. I used a gradient color to represent the continuous and with red color, it can give a sense of danger so darker red represents high PM 2.5 value which is dangerous. After categorizing air quality levels, most of Utrecht area has good air

quality. The north of Utrecht and Centrum area has the worst air quality. To find areas that have the best air quality, I applied another criteria as defined in Table 9.

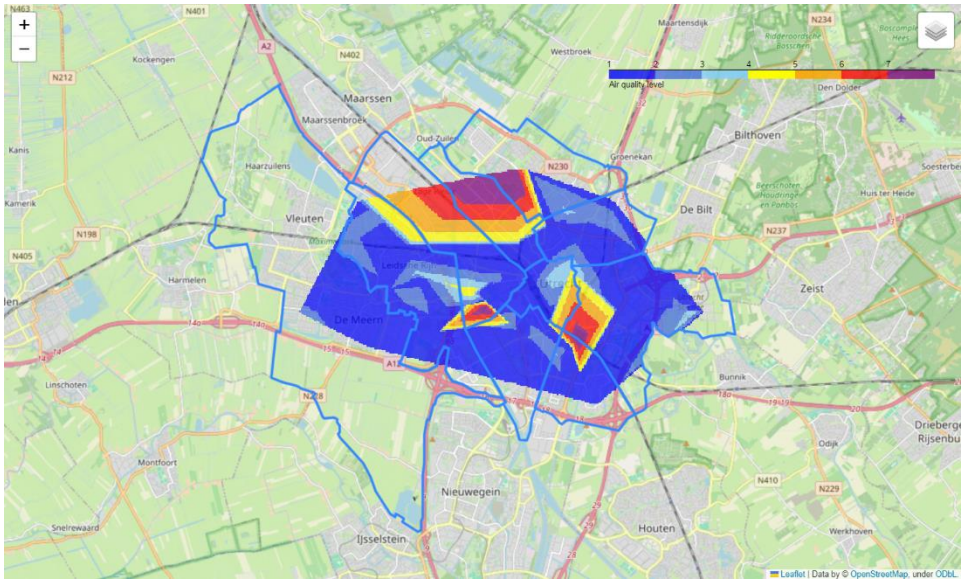

Figure 27 Air quality level based on criteria in Table 9

Url to map in Figure 27: https://gisedu.itc.utwente.nl/student/s3234223/airquality-analysis/map/pm25_categorized_airqual.html

Figure 27 shows the areas that have the best air quality throughout the year (dark blue area) regarding to PM 2.5 value which covers at least some part of all subareas in Utrecht.

NO_2

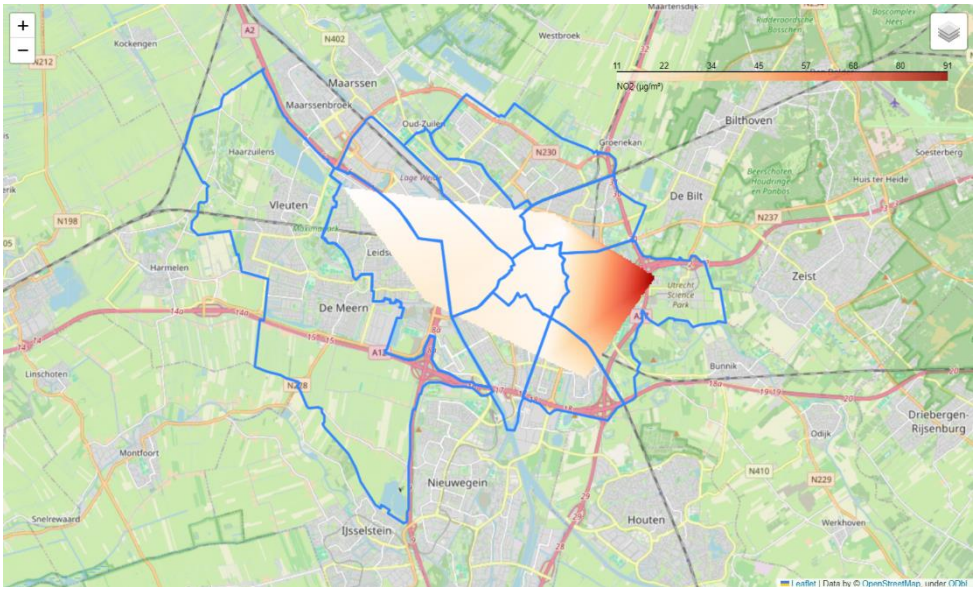

Figure 28 Annual average of NO_2 value

Url to map in Figure 28: https://gisedu.itc.utwente.nl/student/s3234223/airquality-analysis/map/no2_annual_average.html

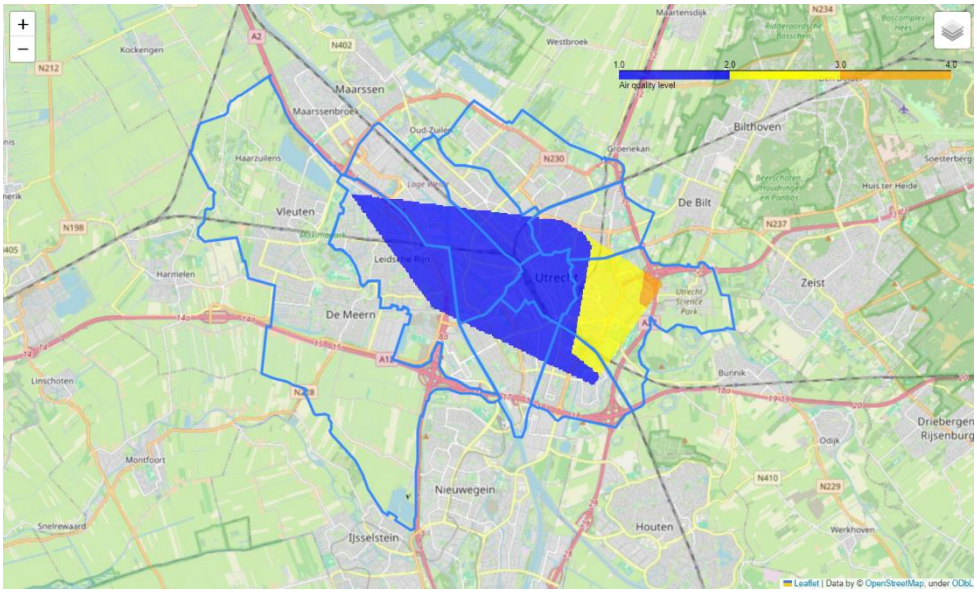

Figure 29 Air quality level based on criteria in Table 8

For NO_2 data, the interpolation result covered fewer areas than PM 2.5 result due to the smaller number of inputs (stations). Most of the covered area also has good air quality and the worst air quality area is on the east side of Utrecht near Utrecht Science Park.

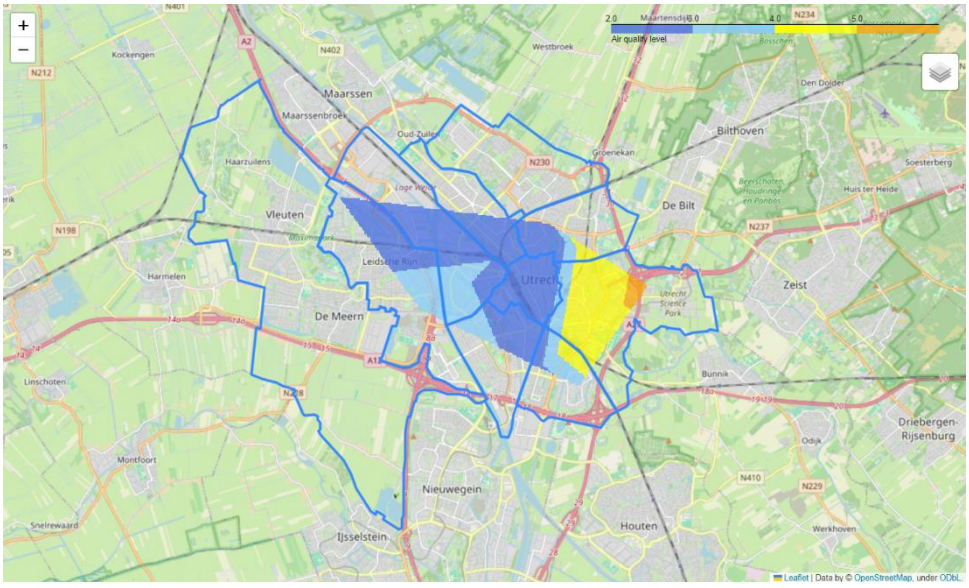

Figure 30 Air quality level based on criteria in Table 9

Url to map in Figure 30: https://gisedu.itc.utwente.nl/student/s3234223/airquality-analysis/map/no2_categorized_airqual.html

After categorizing based on the criteria in Table 9, Figure 30 shows the areas that have the best air quality in darker blue color. Most of Wijk 06 Binnenstad area including Utrecht Central Station has the best air quality (categorized by using NO_2 value).

3. Which areas of the city have the most and least reliable air pollution data?

PM 2.5

Spatial distribution of PM 2.5 stations (from Samenmeten and Luvhtmeetnet):

https://gisedu.itc.utwente.nl/student/s3234223/airquality-analysis/map/pm25_stations_distribution.html

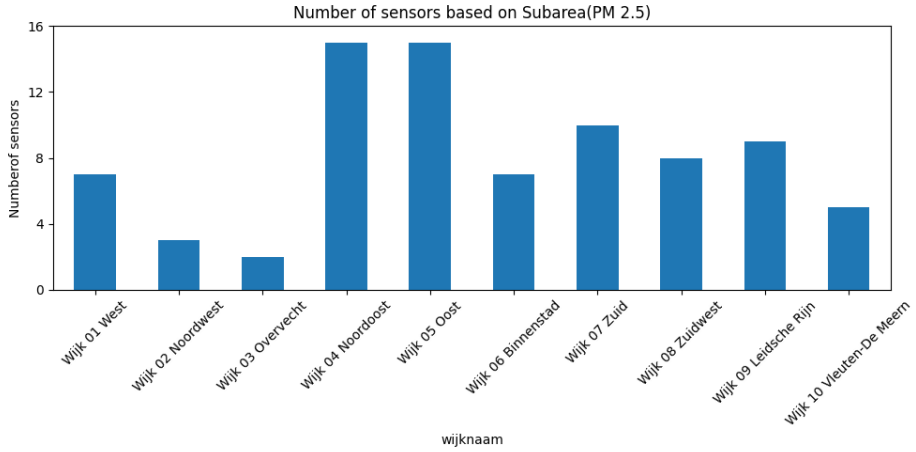

Figure 31 Number of Stations based on Subarea (PM 2.5 sensors)

From Figure 31, two subareas have the highest number of stations, 15 stations, which are Wijk 04 Noordoost and Wijk 05 Oost. Wijk 03 Overvecht has the least number of stations which is 2 stations.

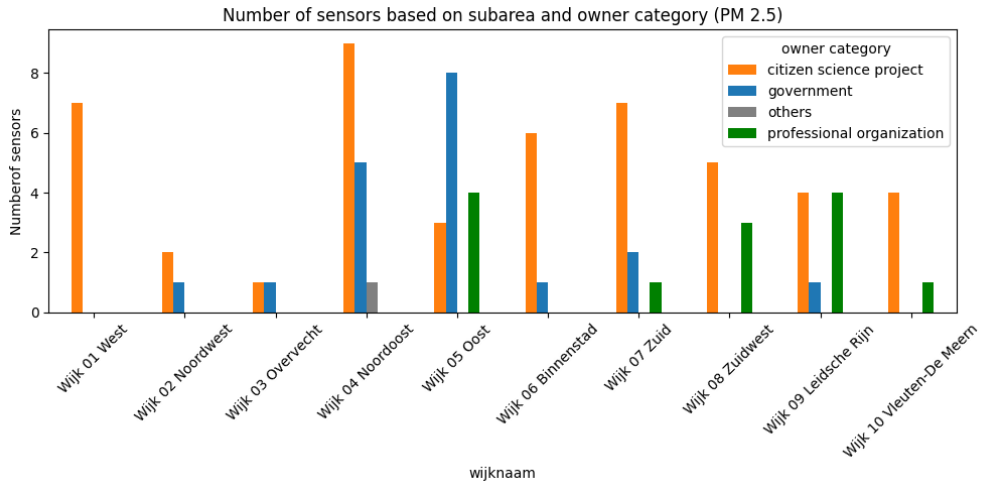

Figure 32 Number of Stations based on subarea and owner category (PM 2.5)

In Wijk 05 Oost, most stations are owned by government, professional organizations, and citizen science project respectively. While most of stations in Wijk 04 Noordoost are owned by citizen science project. I would say that now Wijk 04 Noordoost and Wijk 05 Oost area have the most reliable PM 2.5 data and the area that has the least reliable PM 2.5 data is Wijk 03 Overvecht. To confirm the decision, sensor types need to take into account as well.

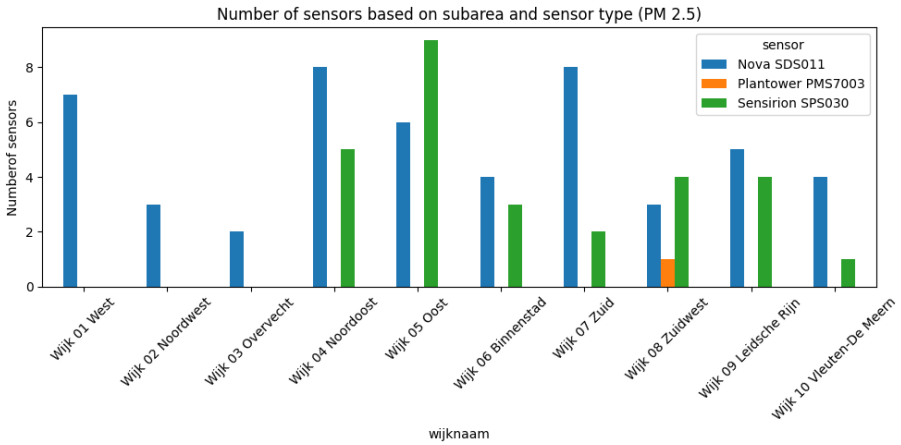

Figure 33 Number of stations based on subarea and sensor type (PM 2.5)

Both Wijk 04 and 05 has same sensor types, SDS011 and SPS030. However, Wijk 05 has higher number of SPS030 sensors which has the highest quality among three sensor types regarding to Table 3.

In conclusion, Wijk 05 Oost has the most reliable PM 2.5 data while Wijk 03 Overvecht has the least reliable PM 2.5 data.

NO_2

Spatial distribution of NO_2 station (from Samenmeten and Luchtmeetnet):

https://gisedu.itc.utwente.nl/student/s3234223/airquality-analysis/map/no2_stations_distribution.html

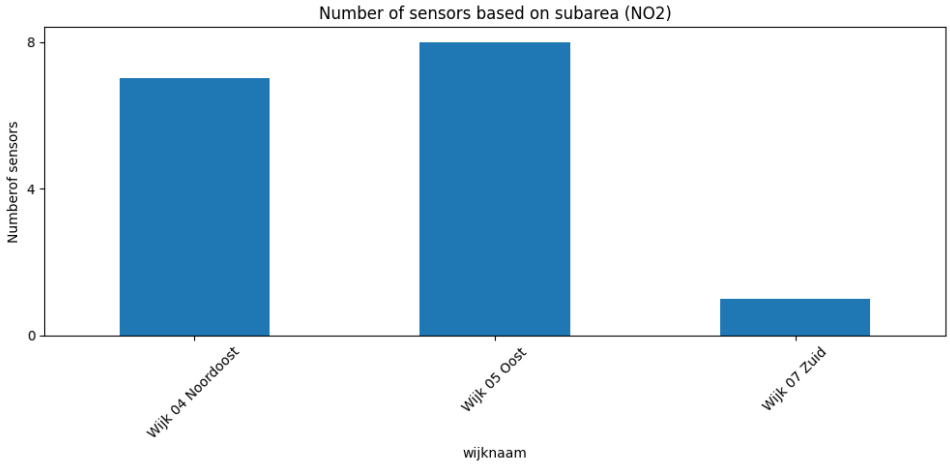

Figure 34 Number of stations based on sub area (NO_2 sensors)

There are only three areas that has NO_2 sensors and the area that has the highest number of NO_2 sensors is Wijk 05 Oost which is 8 sensors/ stations.

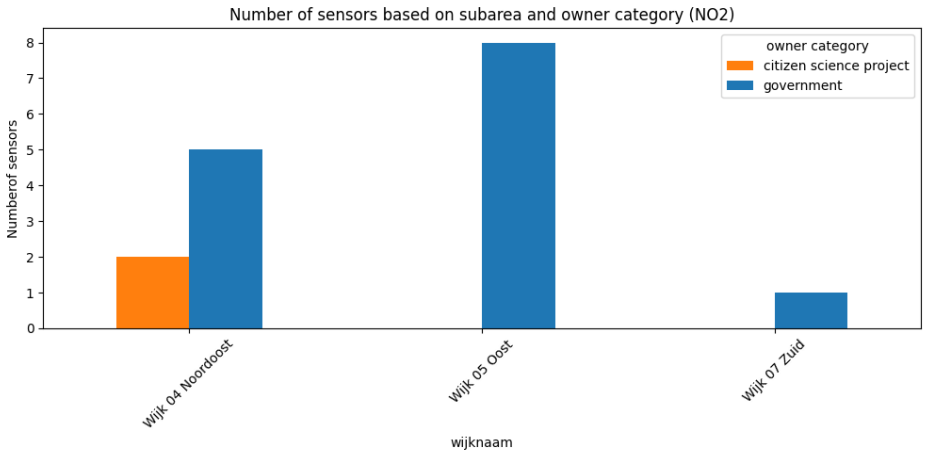

Figure 35 Number of stations based on owner category

From Figure 35, it shows that all stations in Wijk 05 are owned by the government so it must be pretty reliable. For NO_2 data, Wijk 05 Oost has the most reliable data and other Wijks except Wijk 04 and 07 have the least reliable data.

Overall, Wijk 05 Oost has the most reliable PM 2.5 and NO_2 data while Wijk 03 Overvecht has the least reliable data for Both PM 2.5 and NO_2 data.

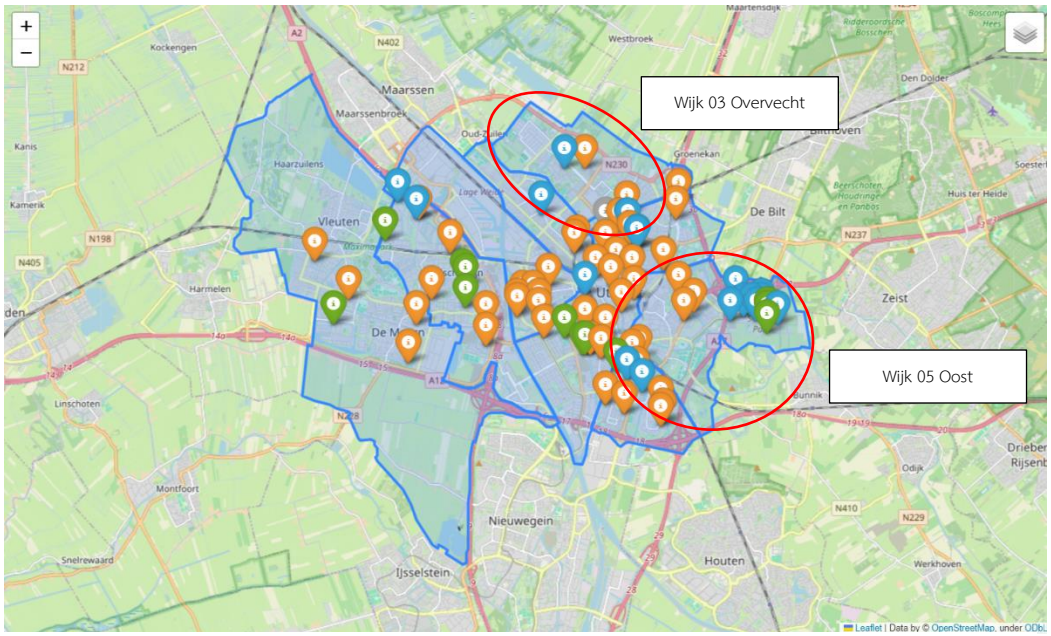

Figure 36 Utrecht subarea reference

Use of AI

In this project, I used ChatGPT to optimize my code to be more efficient in some parts, used it for finding the plot function such as correlation heatmap, and learned more about available interpolation method and application of them in python as well. For the report itself, I used Grammarly to check the grammar and spelling.

References

European Union. (2019, August 13). *Sniffer Bike - a project to track air quality in Utrecht.* https://data.europa.eu/en/news-events/news/sniffer-bike-project-track-air-quality-utrecht

RIVM. (2023, July 12). *Luchtkwaliteitsdata als CSV bestanden*. https://data.rivm.nl/data/luchtmeetnet/readme.pdf

RIVM. (n.d.). *Veelgestelde vragen dataportaal.* https://www.samenmeten.nl/dataportaal/veelgestelde-vragen-dataportaal

RIVM. (n.d.). *Sensoren voor fijnstof (PM2,5/PM10).* https://www.samenmeten.nl/sensoren-voor-fijn-stof-pm25pm10

Sensirion. (2020, March). *Datasheet SPS30*. https://sensirion.com/media/documents/8600FF88/616542B5/Sensirion_PM_Sensors_Datasheet_SPS30.pdf

Zhou. (2016, June 1). *PMS7003 series data manual*. https://aqicn.org/air/sensor/spec/pms7003-english-v2.5.pdf

Luchtmeetnet. (n.d.). *Negatieve waarden*. https://www.luchtmeetnet.nl/informatie/overige/negatieve-waarden

Jaadi. (2019, October 15). *Everything you need to know about interpreting correlations*. https://towardsdatascience.com/eveything-you-need-to-know-about-interpreting-correlations-2c485841c0b8

NS. (n.d.). *Travelling by train with NS.* https://www.ns.nl/binaries/_ht_1449754213072/content/assets/ns-en/22471_nsr_brochure_travelling_by_train_a5.pdf

Luchtmeetnet. (n.d.). *Airquality index (LKI).* https://www.luchtmeetnet.nl/informatie/luchtkwaliteit/luchtkwaliteitsindex-(lki)