Air Quality Project

La Salle Campus Barcelona: Data Analysis Tools 2022/2023

Hugo González Grañena

05/23

Contents

BACKGROUND

One of the main concerns of the citizens in Barcelona is the pollution in the city, which is why we have decided to analyse the quality of the air. In the following document you will find a concrete analysis of the pollution levels, and the different factors impacting the quality of the air.

The report provides an extensive and comprehensive analysis of our work and the overall conclusions.

DATASETS DESCRIPTION

Sensors Dataset

Barcelona has a number of sensors distributed among the city. These sensors are constantly (hour by hour) measuring the quality of the air.

The information about these sensors and its locations can be found by following this link.

By clicking on "View complementary description and field definition" you will get further information.

Measurements Dataset

The measurements that the sensors are doing can be found here.

As you can see we have a different file for each month and later in the project we will transform the datasets to put all information of 2022 together.

Another mandatory step is to understand the data provided. We put a lot of emphasis on analysing the data, refering to the documentation available on the website.

Contaminants Dataset

Several contaminating polluting factors are being analyzed in the report, their description being found here.

Trees Dataset

There is something else we want to consider. For each of the sensors deployed, we also want to find out if there is any relationship between the pollution levels detected in the neighborhood (area) they are located and the number of trees in that neighborhood (area).

The trees in green areas can be found here.

The trees on the streets can be found here.

Finally, the size of the neighborhood (area) matters because the density is a critical factor here. The surface of each area, as well as the population, can be found here.

DATASET NAMES TRANSLATIONS

All the data acquired were in Catalan, so for a better understanding of each part of the data, we have translated them.

QUALITAT AIRE DETALL

FIELD	DESCRIPTION
CODI_PROVINCIA	PROVINCE CODE
PROVINCIA	PROVINCE
CODI_MUNICIPI	MUNICIPAL CODE
MUNICIPI	MUNICIPALITY
ESTACIO	STATION
CODI_CONTAMINANT	CONTAMINANT CODE
ANY	YEAR
MES	MONTH
DIA	DAY
H...	HOURS
V...	VALIDATION

QUALITAT AIRE ESTACIOS

FIELD	DESCRIPTION
ESTACIO	STATION
NOM_CABINA	STATION/ CABIN NAME
CODI_DTES	STATION/ CABIN CODE
ZQA	AIR CODE QUALITY,WHERE THE STATION,LOCATION
CODI_EOI	EUROPEAN CODE STATION/ CABIN
LONGITUD	LONGITUDE
LATITUD	LONGITUDE
UBICACIO	ADRESS/ CROSSROADS
CODI_DISTRICTE	DISTRICT CODE
NOM_DISTRICTE	DISTRICT NAM
CODI_BARRI	NEIGHBORHOOD CODE
NOM_BARRI	NEIGHBORHOOD NAME
CLAS_1	TYPE OF STATION
CLAS_2	TYPE OF STATION
CODI_CONTAMINANT	CONTAMINANT CODE

SENSOR MEASUREMENTS ALL MONTHS

FIELD	DESCRIPTION
CODI_PROVINCIA	PROVINCE CODE
PROVINCIA	PROVINCE
CODI_MUNICIPI	MUNICIPAL CODE
MUNICIPI	MUNICIPALITY
ESTACIO	STATION

FIELD	DESCRIPTION
CODI_CONTAMINANT	CONTAMINANT CODE
ANY	YEAR
MES	MONTH
DIA	DAY
H...	HOURS
V...	VALIDATION

QUALITAT AIRE CONTAMINENTS

FIELD	DESCRIPTION
CODI_CONTAMINANT	CONTAMINANT CODE
DESC_CONTAMINANT	CONTAMINANT DESCRIPTION
UNITATS	CONTAMINANT MEASUREMENTS UNITS

STREET TREES

FIELD	DESCRIPTION
CODI	ITEM CODE
X_ETRS89	STH X COORDINATE
Y_ETRS89	STH Y COORDINATE
LATITUD	LATITUDE
LONGITUD	LONGITUDE
TIPUS_ELEMENT	ITEM TYPE
ESPAI_VERD	GREEN SPACE TO WHICH THE TREE BELONGS
ADRECA	ADRESS
CAT_ESPECIE_ID	TREE SPECIES ID
NOM_CASTELLA	SCIENTIFIC NAME OF THE TREE
NOM_CASTALLA	SPANISH NAME OF THE TREE
NOM_CAT	CATALAN NAME OF THE TREE
CATEGORIA_ARBRAT	COTHEGORY OF THE TREE
DATA_PLANTACIO	DATE OF PLANTING
TIPUS_AIGUA	TYPE OF IRRIGATION WATER
TIPUS_REG	TYPE OF THE IRRIGATION REQUIRED
GEOM	LOCATION GEOMATRY
CATALOGACIO	CATALOGING
CODI_BARRI	NEIGHBORHOOD CODE
NOM_BARRI	NEIGHBORHOOD NAME
CODI_DISTRICTE	DISTRICT CODE
NOM_DISTRICTE	DISTRICT NAME

FIRST PART OF THE PROJECT

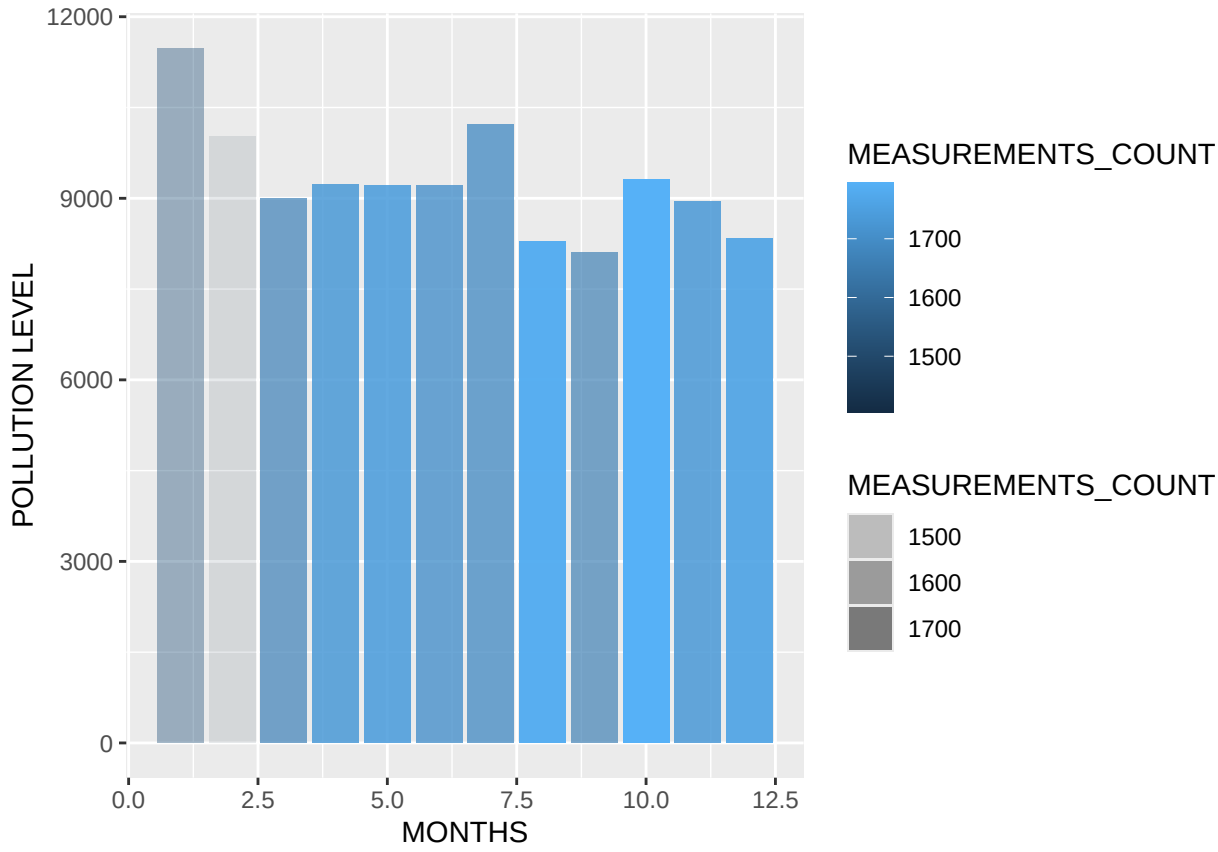
DATA PREPARARTION

Firstly, we need to read the data sets.

For the data set describing the measurements that the sensors took for January to December, we read them seperatly, then unite them by using the `union()` function.

We identify a lot of NAs, so we eliminate them by using the `na.omit()` function.

THE AVERAGE POLLUTION LEVELS OF EACH MONTH



COMMENTS

This plot uses the data from the "pollution_months" data frame and sets the x and y axes based on the "MES" and "mean_pollution" columns, respectively. In addition, it uses the column "number" to determine the color and transparency of the bars. The plot is represented using the geom_bar function in a "dodge" position, which indicates that the bars are overlapped but shifted to show a better comparison. Therefore, this plot represents the average pollution data as a function of month and uses different shades to indicate the number of observations in each month.

We can see that the highest point starts in the first month, January, reaching over 11,000 of contamination.

On the other hand, we see that the lowest point of contamination is found at 8000 in the ninth month, September.

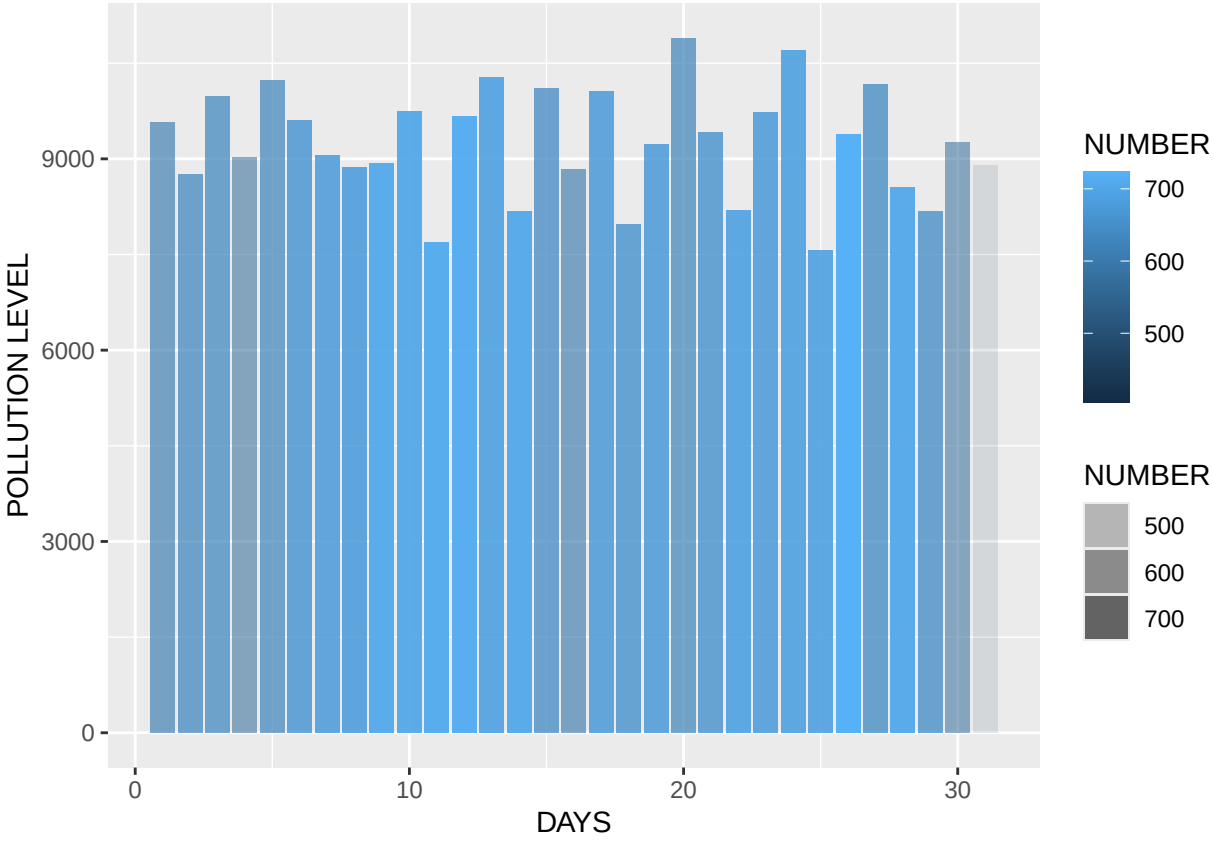
Similarly, if we look at the path of the median of each month we see that we start with a peak in month 1, as we said before.

Then we go down until we reach one of the low points at 9000 pollution around month 4 (April) and we have another peak, not as high as the first one, but significantly high in month 6 (June) around 9500 pollution.

Then, we go down again a little bit reaching the low point in month 9 (September) and finally around month 12 (December) we go up again closing the year in about 9000 of pollution.

With this plot we can see reflected how air pollution progresses during the different months of the year 2022 starting with a peak and progressively decreasing.

THE AVERAGE POLLUTION LEVELS OF EACH DAY OF THE MONTH



COMMENTS

This plot is a bar chart. The data being plotted is from a data frame called "pollution_days" and the x-axis is labeled "DAYS" and is mapped to the column "DIA" in the data frame. The y-axis is labeled "POLLUTION LEVEL" and is mapped to the column "mean_pollution" in the data frame. The fill and alpha aesthetic are mapped to the column "number" in the data frame, which likely represents the number of observations or a count. The position of the bars is set to "dodge", meaning that the bars are separated out so that they don't overlap.

Therefore, this plot represents the average data of pollution as a function of the day and uses different shades to indicate the number of observations on each day. Additionally, it is shown in a bar format that allows to identify a relative comparison of the data for each day of the month.

The highest point is found on the 20th day, being the pollution above 11,000.

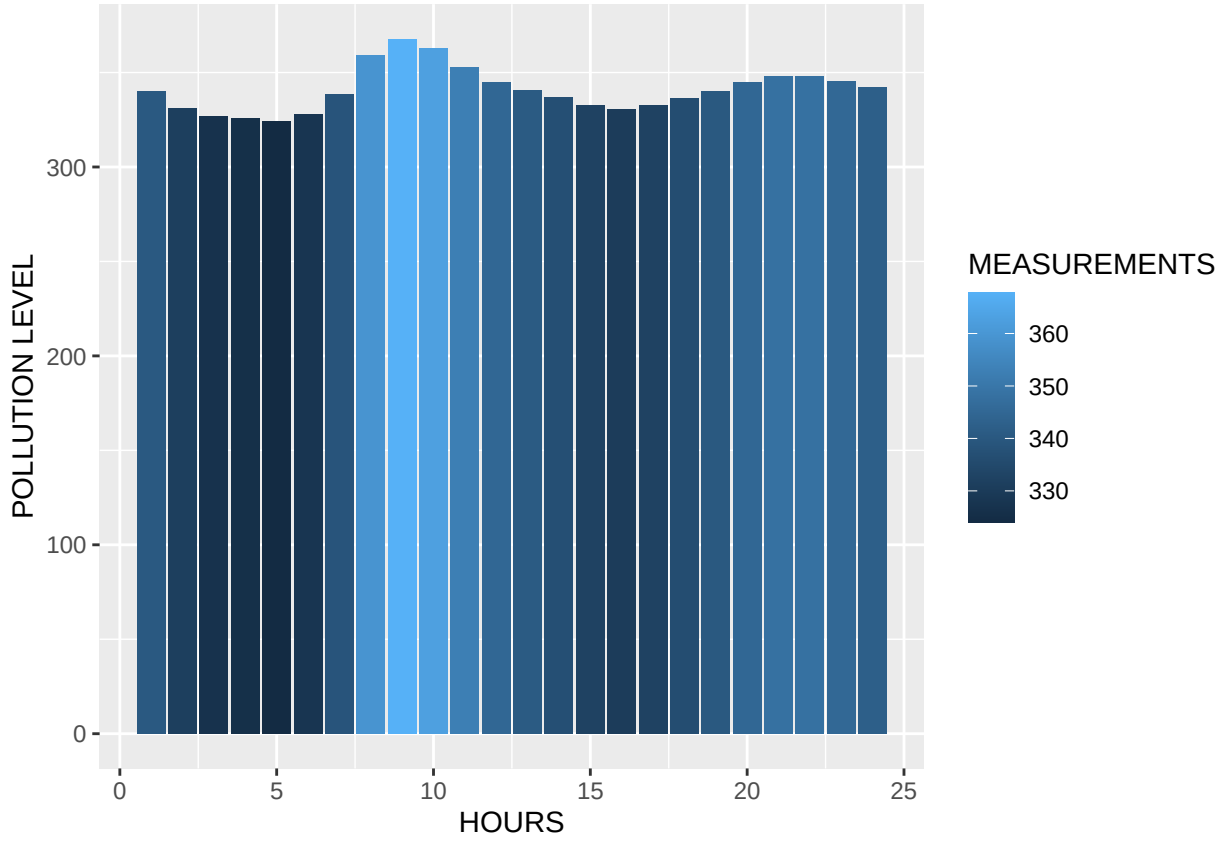
On the other hand the lowest point we find it on day 25th at a contamination level below 8000.

A curious fact is that after day 30, we find a bar which is almost transparent which would be the day 31. This is much transparent because not all months have 31 days and it is below 9000 of contamination.

Analyzing this plot, we realize that during all days, the pollution levels are very diverse, although the first 10 days are usually contained in a fairly centralized between 8500 and 10,250 pollution.

We also note that after day 10th hey expand much more as the lowest point is on day 11 and 25 and the highest point on day 20 and 24.

THE AVERAGE POLLUTION LEVELS OF EACH HOUR OF THE DAY



COMMENTS

This plot is also a bar chart. The data being plotted is from a data frame called “pollution_hours” and the x-axis is labeled “HOURS” and is mapped to the column “hours” in the data frame. The y-axis is labeled “POLLUTION LEVEL” and is mapped to the column “measurements” in the data frame. The fill aesthetic is mapped to the “measurements” column, which is the same as y-axis, this suggest that the bars are colored based on their height. The position of the bars is set to “dodge”, meaning that the bars are separated out so that they don’t overlap. This plot shows the distribution of pollution levels across different hours of the day.

In terms of progression, we can see that the first bar starts at 340 pollution at 01:00. Progressively, this pollution is reduced, reaching its lowest point at 5:00 a.m. around 329 pollution. Then we see that it rises

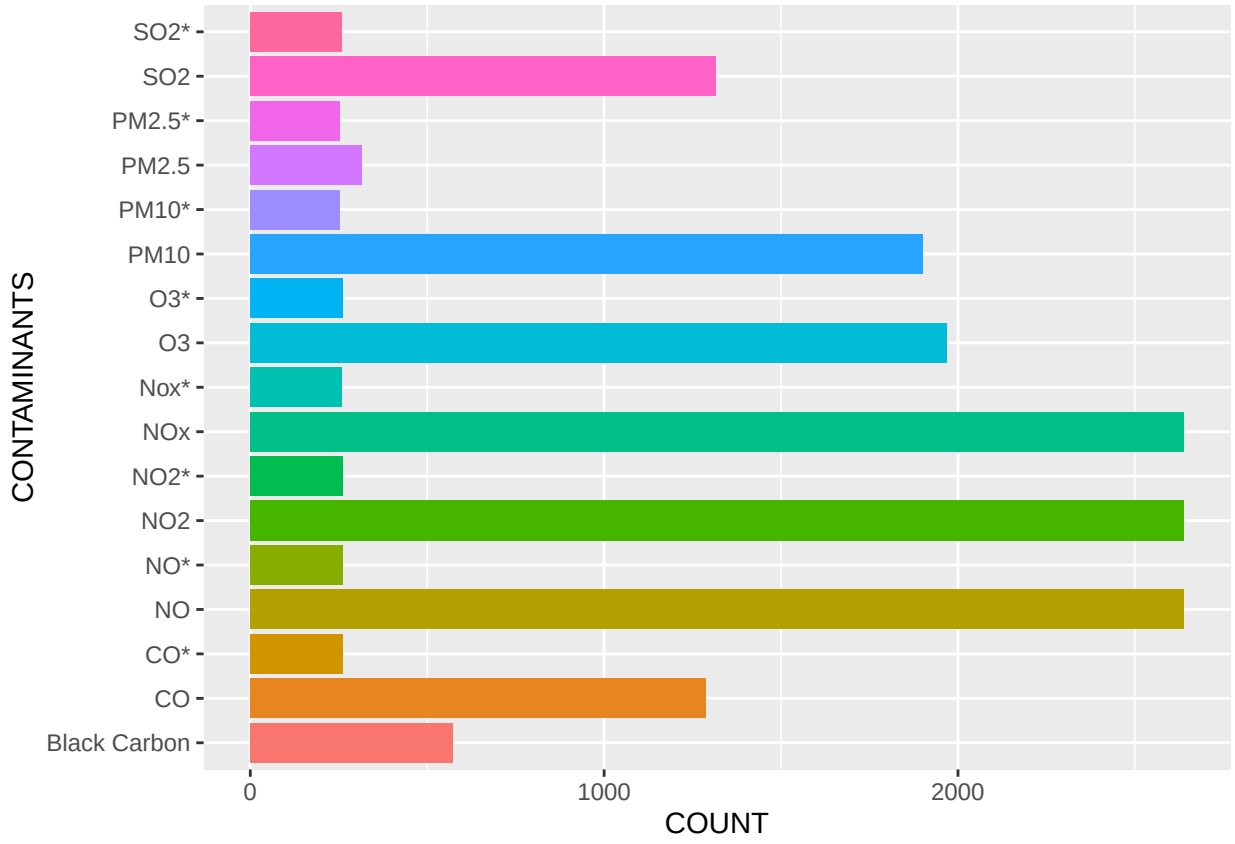
exaggeratedly until reaching its highest point at 9:00 in the morning reaching 372 of pollution approximately. Once it reaches its highest point, it decreases again to 332 at 16:00 hours. Finally, it gradually increases again to 350 pollution at around 9:10 pm. Finally, it decreases a little, ending at 24:00h at 345 pollution.

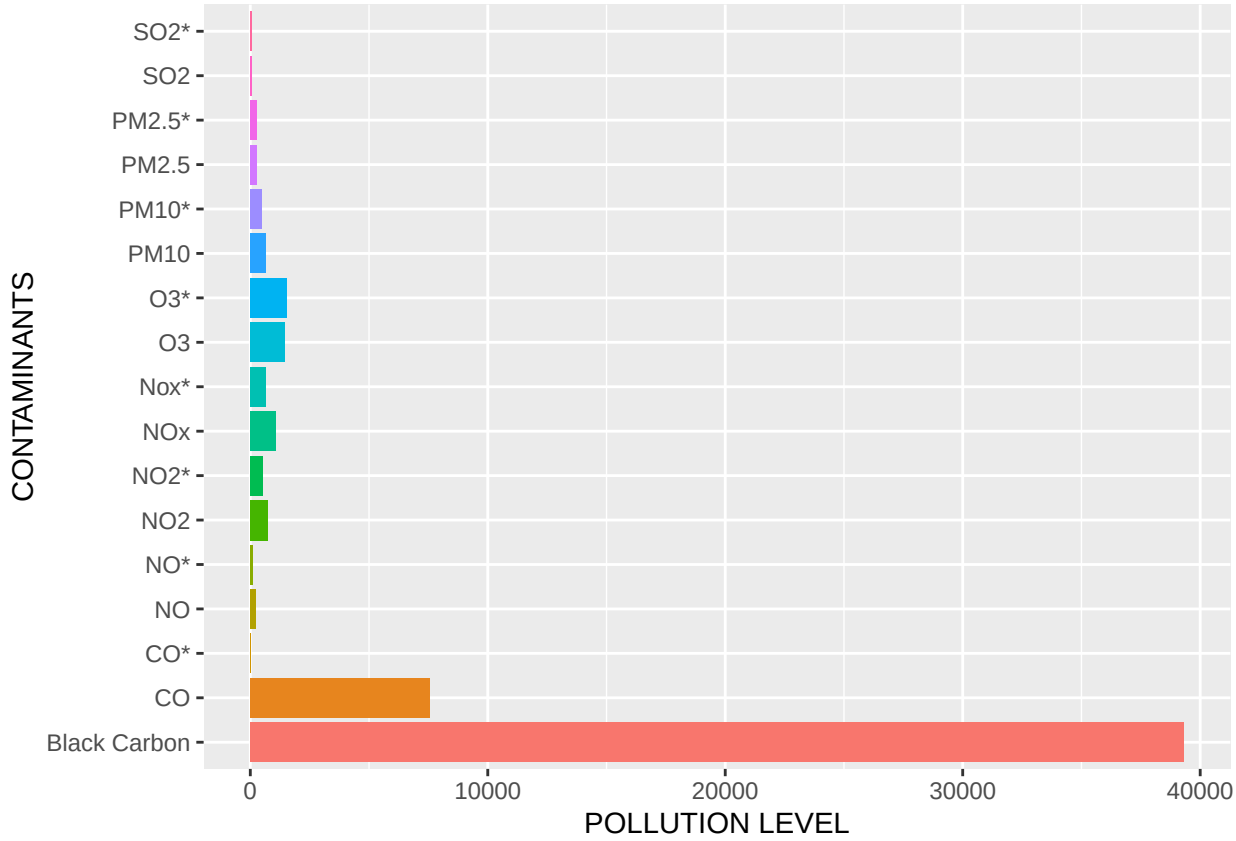
We start at 330 of pollution at 1:00 in the morning. It continually rises until it peaks at 358 around 9:00 am and then falls to 338 between 15:00 and 18:00. Finally it rises again to end at 24:00h at 351 pollution.

With this plot, we can find possible patterns of pollution and hours, it can be used to observe if the pollution levels have a trend by the hour of the day, for instance high pollution levels in the morning, lower during the afternoon and high again during the night.

TYPE OF CONTAMINANT - POLLUTION RELATIONSHIP

We observe that the pollution factor CO is not measured in the same unit as the other ones. Therefore, we transform the pollution level measured by multiplying it by 1000. Then, we plot the results





COMMENTS

In this case, we have 2 plots.

In the plot located in the left, we see a bar chart representing the distribution of different types of contaminants. The data being plotted is from a data frame called "sensor_measurements" and the y-axis is labeled "CONTAMINANTS" and is mapped to the column "Desc_Contaminant" (air pollutants) in the data frame. The x-axis is labeled "COUNT" (amount of pollution of each one of them). The fill color of the bars is determined by the value of the "Desc_Contaminant" variable, this suggest that the bars are colored based on different contaminants. This chart is showing the count of contaminants grouped by type, with bars in descending order according to the frequency of the contaminant. The option show.legend = FALSE removes the legend from the plot. This chart provide an insight of the most common contaminants that the sensor detect and shows the distribution of different contaminants types in the sensor_measurements dataset.

If we analyze this graph, we observe that the three most polluting elements in the air are nitrogen oxides (NOX), NO2 and nitric oxide (NO) reaching almost 2500 µg/m³. These are atmospheric pollutants whose main sources are road traffic as well as emissions from certain industries and coal-fired heating systems (no longer in use).

And then we have ozone (O3) which is further behind, passing 1750 µg/m³ of pollution.

Then we have PM10 which does not reach 1750 µg/m³ of pollution. PM10 is defined as small solid or liquid particles of dust, ash, soot, metallic particles, cement or pollen, dispersed in the atmosphere, and whose aerodynamic diameter is less than 10 µm.

Almost at the same level of pollution, at 1200µg/m³, we have sulfur dioxide (SO2) and carbon monoxide (CO) that is found in gases produced by boilers, kerosene heaters, vehicles whose engines are "warmed up" in garages, stoves, portable lamps or lanterns, gas stoves, portable generators, or the burning of coal or wood.

Finally, we have Black Carbon and Biomassa Black Carbon at just over 500 µg/m³ of pollution, and PM2.5 at just over 250 µg/m³ of pollution.

If we move to the plot located in the right side, we can see that this plot, in terms of aesthetics, is similar to the one commented before.

The data being plotted is from a data frame called "pollution_factors". The y-axis is labeled "CONTAMINANTS" and is mapped to the column "Desc_Contaminant" in the data frame. The x-axis is labeled "POLLUTION LEVEL" and is mapped to the column "pollution_levels" in the data frame. The fill aesthetic is also mapped to "Desc_Contaminant" as in the plot before, this suggest that the columns are colored based on different contaminants. The geom_col function is used to create the columns, which represents the pollution level of each contaminant. The show.legend is set to FALSE, this means that the legend will not appear in the plot. This plot shows the distribution of pollution levels of different contaminants types in the pollution_factors dataset.

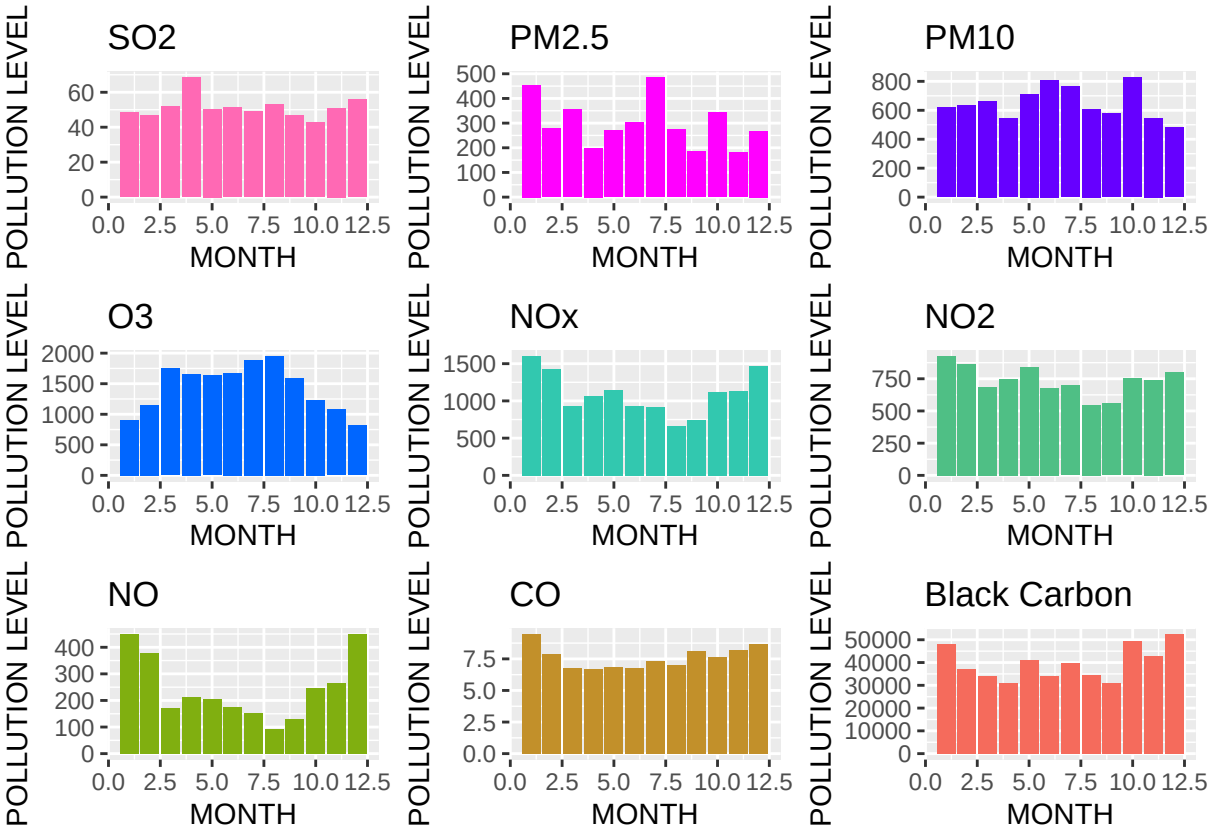
Analyzing the results of this plot, it can observed that the contaminant that has the highest pollution level is the Black Carbon arriving almost to the 40.000.

Then we have CO around the 7500 of pollution level creating a big difference of level compared to the Black Carbon contaminant.

Finally, we find the rest of the pollutants below the 2500 pollution level.

If we organize these pollutants in order of most pollutants at least the ranking would be as follows: 1. Black Carbon 2. CO 3. O3 4. Nox 5. NO2 6. PM10 7. PM2.5 8. NO 9. SO2

SHOW FOR EACH CONTAMINANT THE AVERANGE POLLUTION LEVELS IN EACH MONTH



COMMENTS

This plot is using the ggplot and patchwork package in R to create 8 column charts. The data being plotted is from several data frames called “measurements_SO2”, “measurements_PM2.5”, “measurements_PM10”, “measurements_O3”, “measurements_NOx”, “measurements_NO2”, “measurements_NO”, “measurements_CO”, “measurements_CO2” respectively, that are the measurements for each air contaminant.

Each plot shows the mean pollution level for each month, for a specific contaminant. Each plot is created by first grouping the data by month, then calculating the mean of the pollution level for that month, and removing missing values. Then, it uses ggplot package to create a column chart with x-axis labeled as “MONTH” and y-axis labeled as “POLLUTION LEVEL” and each plot is filled with different colors, the fill colors are set with a specific color code or a color name in order to match with the contaminant colour set in the plot before.

Finally, patchwork package is used to combine all the 8 plots into one, with different titles for each plot. This plot is showing a comparison of the average pollution levels of different contaminants over different months.

Analyzing this Plot we can see that very few pollutants are constant during all months of the year, that is to say that they maintain the same level of pollution all year round.

In these cases we could say that the most constant pollutants are SO2 and CO. Even so, in the case of SO2 we observe that in month 4 (April), there is a high outlier. And in the case of CO, we see the high outlier in the first month, January, and then the levels decrease progressively until the seventh month (July) when they also increase progressively.

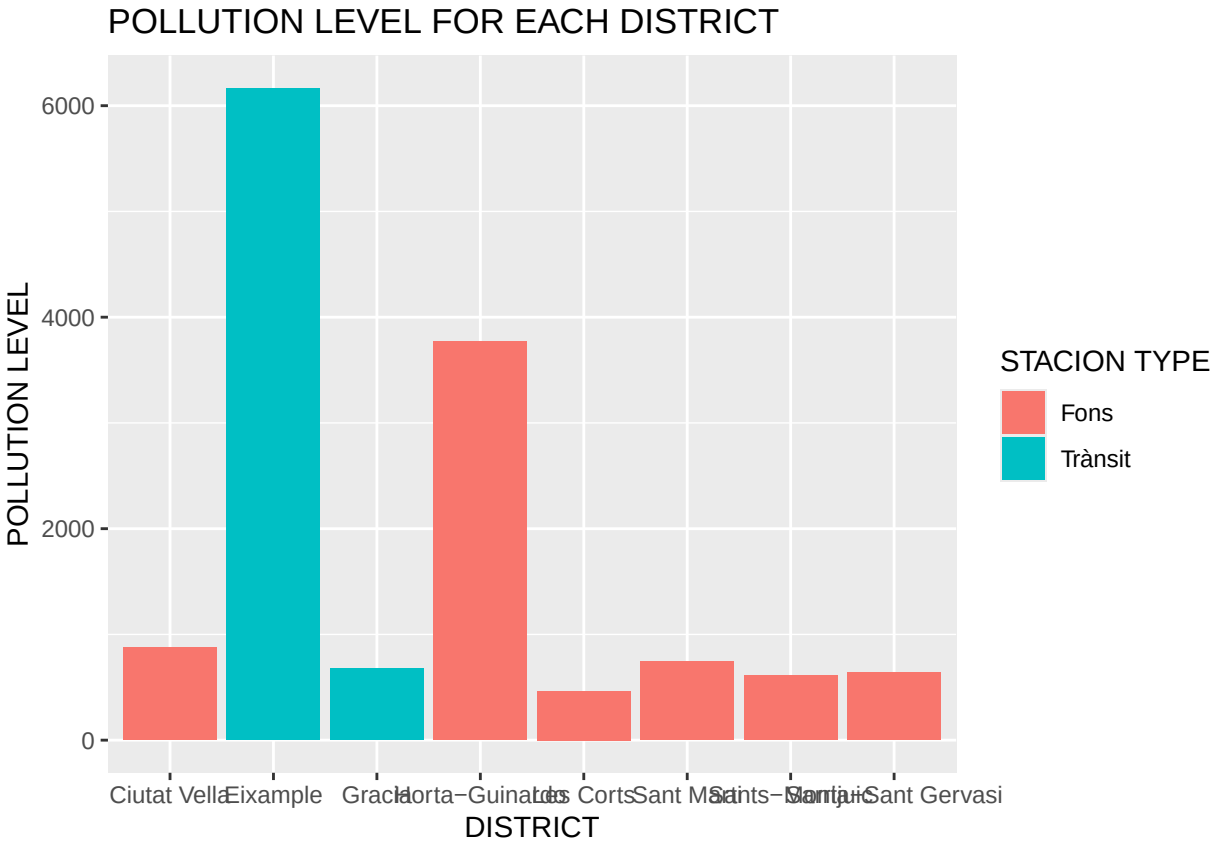
For the other pollutants, different patterns are observed; The pollutant PM2.5 is very irregular since the levels vary greatly each month.

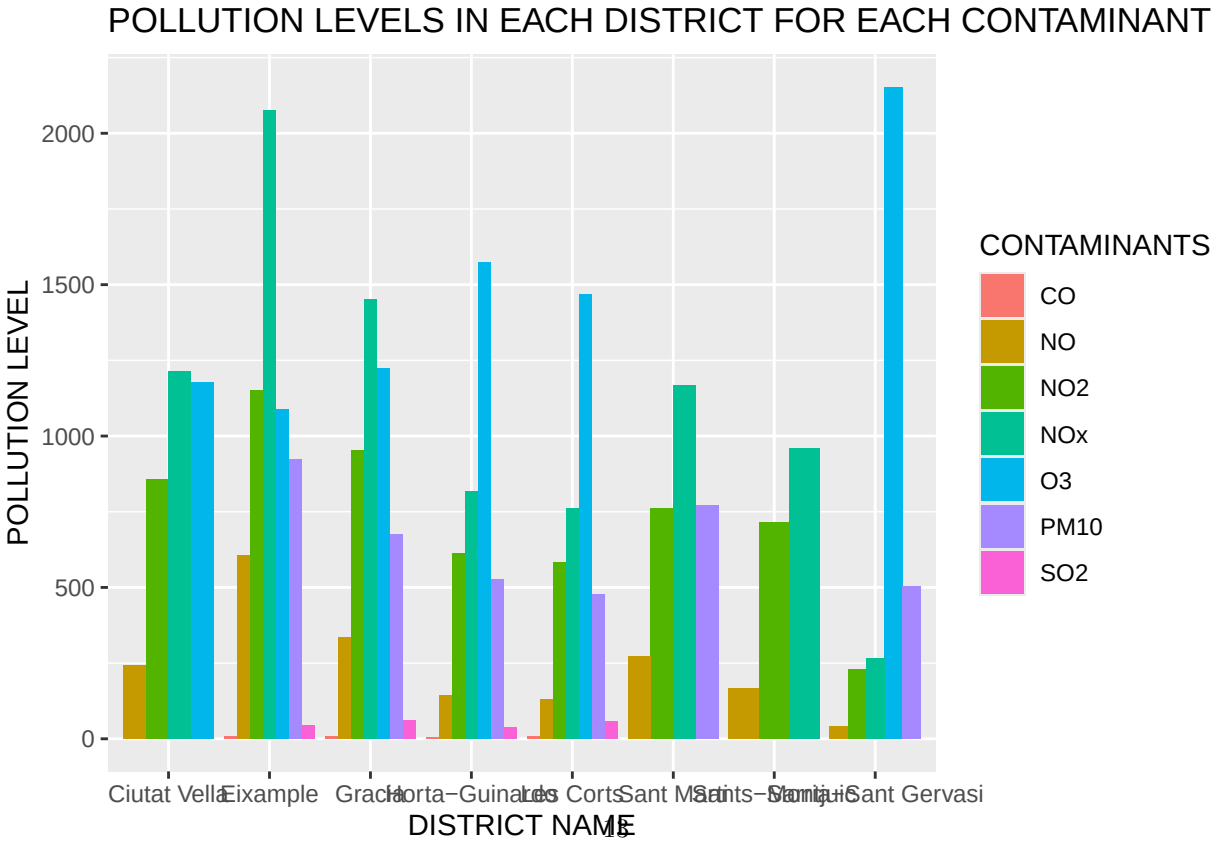
In the case of PM10, the irregularity is not so great and we could classify it as a fairly symmetrical plot since it starts increasing the levels progressively until it reaches a point in the month of June (middle of the year) where it decreases progressively again. In this way a triangle shape is created even if in the month of October we find a higher level compared to its next months (August, September, November or December).

In O3, the levels start and end quite low (below 1000) and between March and September, the levels are higher, staying between 1500 and 2000. In Nox and NO, we can see the opposite, i.e. the levels start and end quite high (above 1500 and 450, respectively) and during the other months of the year they are reduced, reaching their minimum in August, and increasing again until the end of the year.

Finally, in the cases of NO2 and Black Carbon, the plots are quite similar. The levels are not very regular, but neither do we find the difference in level that we had in the previous cases. In the case of NO2 the levels oscillate between 500 and 750 of contamination and in the case of Black Carbon they oscillate between 30,000 and over 50,000. Throughout the year we observe changes from one month to another, reaching peaks at the beginning (NO2) and end (Black Carbon) of the year.

SHOW FOR EACH DITRICT THE POLLUTION LEVELS FOR EACH CONTAMINANT





COMMENTS

The first plot is using the ggplot package in R to create a bar chart. The data being plotted is from a data frame called "sensor_measurements" and it is joined with another data frame "sensor_info_loc". The data is grouped by the column "ESTACIO" and it calculates the average pollution level by summing the columns H01 to H24 and removing missing values. The x-axis is labeled "Nom_districte" which is a column from sensor_info_loc data frame and the y-axis is labeled "POLLUTION LEVEL". The fill aesthetic is mapped to the "Clas_2" column, which likely represents if it's fons or trànsit. The position of the bars is set to "dodge", meaning that the bars are separated out so that they don't overlap. This plot shows the average pollution level by district, where the bars are colored based on the Clas_2 variable.

In this case we see that only two districts have traffic pollution; l'Eixample and Gràcia. In terms of levels, most of them do not reach 1000 of pollution, although l'Eixample exceeds 8000 and Horta-Guinardó almost reaches 4000, being clearly the districts with the highest level of pollution.

The second plot is using the data frame "stations_pollution" (created by filtering the sensor_measurements data frame and joining it with sensor_info_loc data frame) and is showing the average pollution level for each district for each contaminant type. The x-axis is labeled as "DISTRICT NAME" and the y-axis is labeled as "POLLUTION LEVEL". The fill aesthetic is used to distinguish between the different contaminant types and is labeled as "CONTAMINANTS". The plot is using geom_bar with "position = 'dodge' " to display the bars side by side. The title of the plot indicates that it is showing the pollution levels in each district for each contaminant. This plot is useful to visualize and compare the pollution levels across districts and contaminant types. In this case, the colour of the contaminants do not match with the plots before but on the right we have the legend.

If we analyze this plot we can see that the pollutants with the highest level of pollution in all districts are NO2, NOx, and O3.

To analyze each district, we will analyze the levels and pollutants from right to left.

In the first district, Ciutat Vella, no pollutant exceeds 1250. We only observe four pollutants of which the one with the lowest level of contamination is NO.

In the next one, l'Eixample, we can observe that in this district all pollutants are present as in the previous one, most of them do not exceed 1250, although NOx exceeds 2000, being one of the highest points in terms of pollution level, and the pollutants with the lowest levels are CO and SO2, respectively.

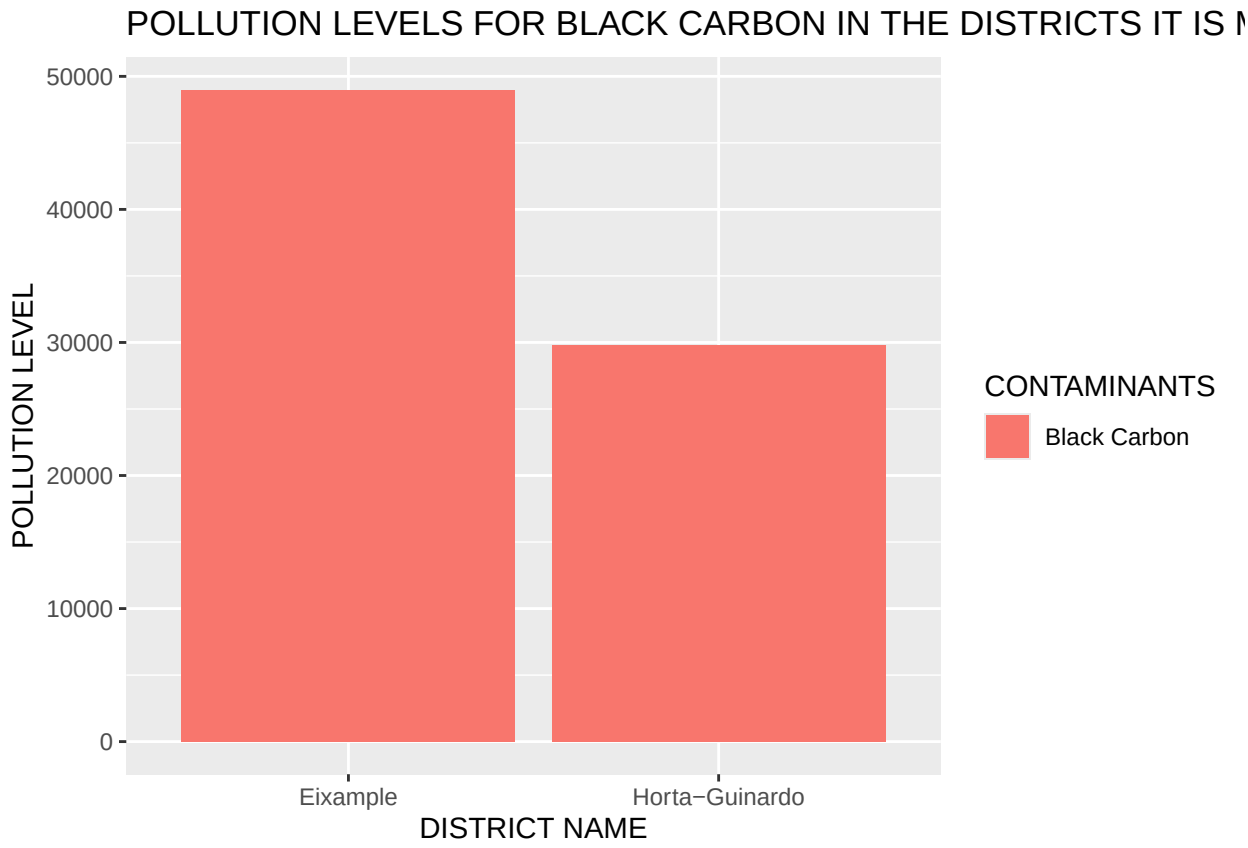
The third, the district of grace also has all the pollutants in this case, as in the previous one, the least pollutant is CO followed by SO2 for less than 100 pollution. If we look at the highest levels of pollution, we have NOx and O3 together with NO2 reaching almost 1500, 1250 and below 1000, respectively.

In Horta-Guinardó (fourth) and les Corts (fifth), the pollutant with the highest level is O3 at around 1500. Next we have NOx, NO2 and PM10 between 500 and 750. And finally we have very low levels of NO, SO2 and CO below 250.

The districts of Sant Martí and Sants-Montjuic have quite similar levels. The main difference is found in the number of pollutants; Sant Martí has Nox, NO2, PM10 and NO while Sants-Montjuic has no PM10.

Finally in the district of Sarrià-Sant Gervasi, we observe that it does not have all the pollutants; it has NO, NO2, Nox, PM10 and O3 of which all are below 500 except O3 which marks the highest peak of the whole plot reaching almost 2250.

POLLUTION LEVELS FOR BLACK CARBON



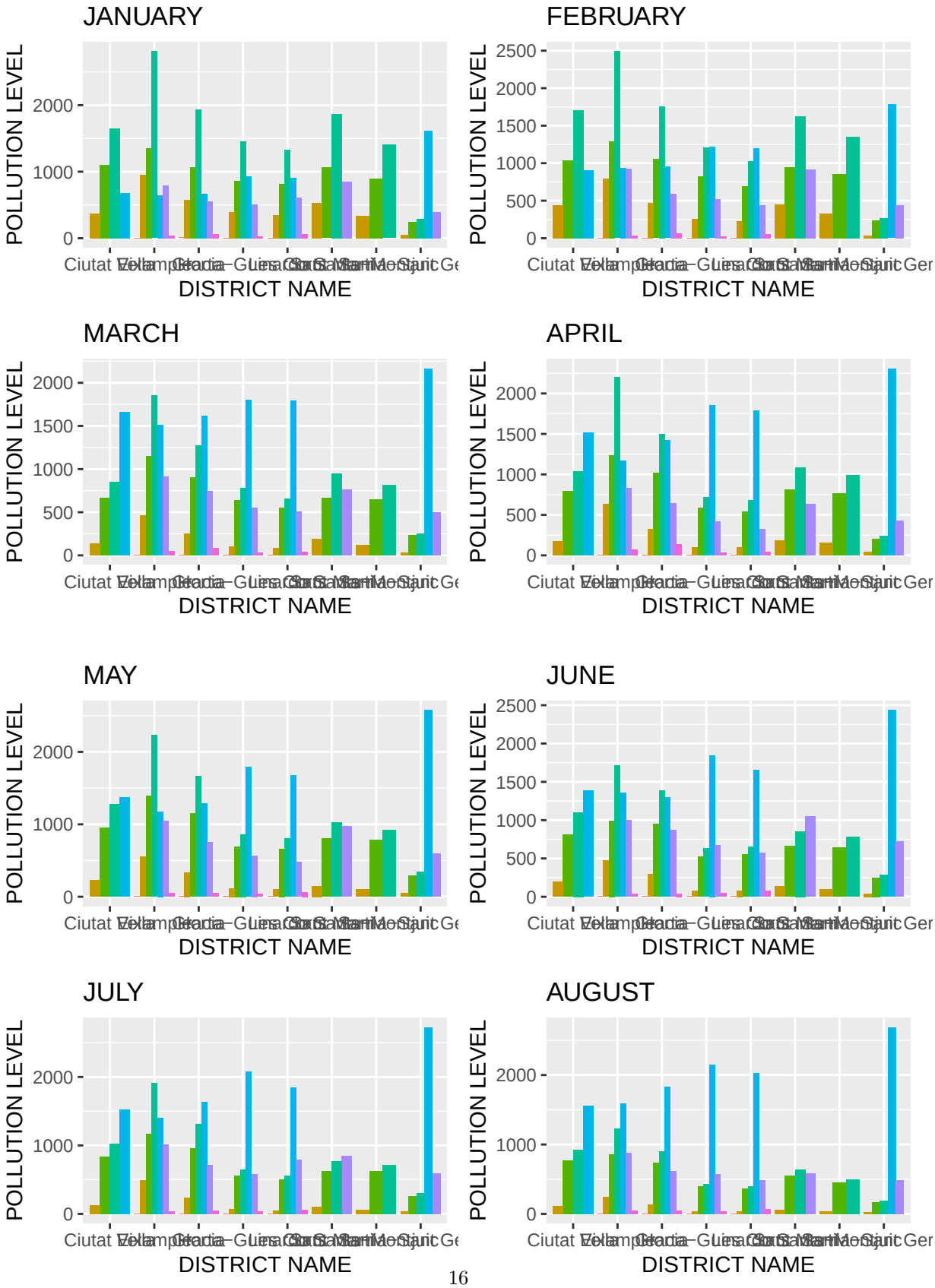
COMMENTS

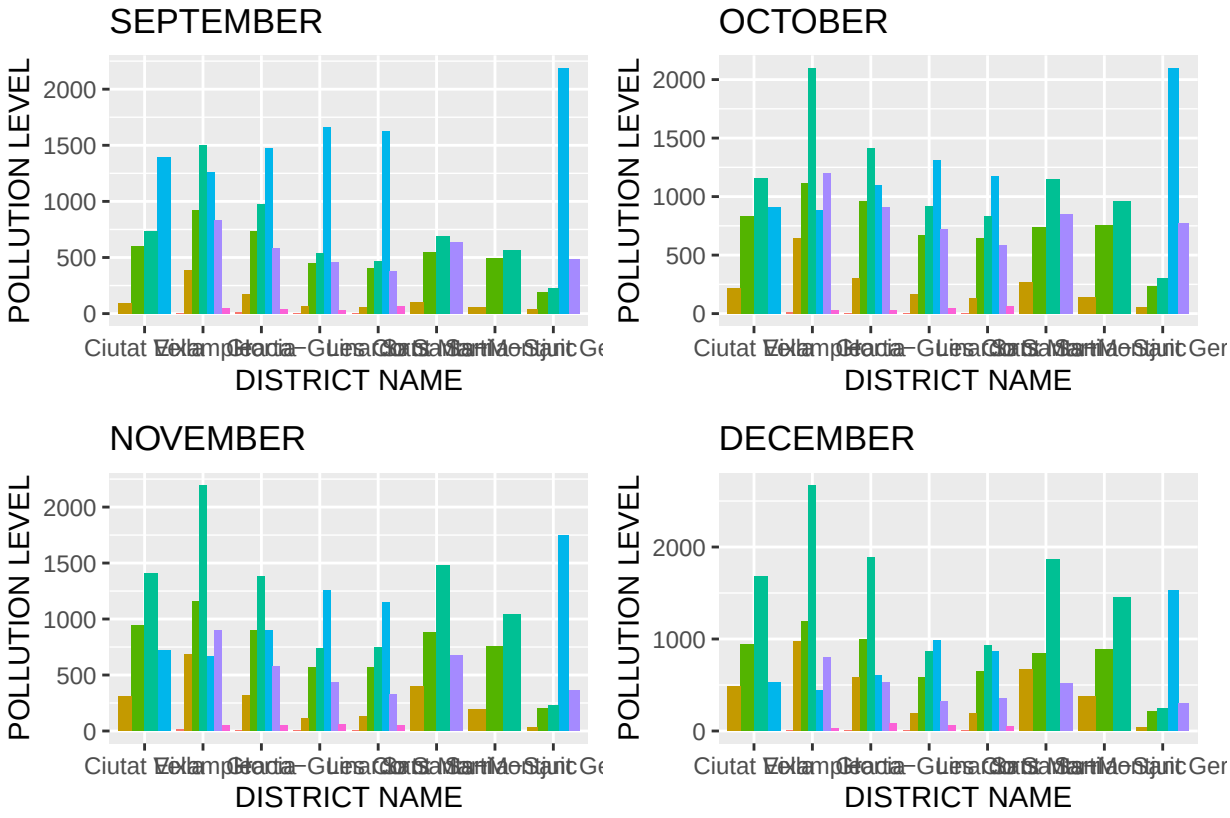
This plot is showing the average pollution level of black carbon in each district, where the measurement is taken. The x-axis is showing the name of the district, the y-axis is showing the pollution level and the fill color is showing the type of contaminant. The data is grouped by district and the mean of the pollution levels is calculated. It's using geom_bar to show the data in a bar chart format with "dodge" position. This plot helps to understand the distribution of black carbon pollution across different districts.

As we can observe, this contaminant is only present in two districts; Eixample and Horta-Guinardó.

L'Eixample has an approximate total of 50000 (polution level)mostly the 100% and Horta a 30000 (polution level), the 60%. At a glance we can see that the area of L'Eixample is much more polluted than Horta, about 40% more. This may be due to the large amount of traffic in these areas.

SHOW FOR EACH DITRICT THE POLLUTION LEVELS FOR EACH CONTAMINANT IN EACH MONTH





COMMENTS

In this plot we can see the presence of gases in each district throughout all months of the year.

We can see that the most common pollutant gas is o3, i.e. ozone, especially in the area of Sant Gervasi. We also see it repeatedly in Ciutat Vella and Horta.

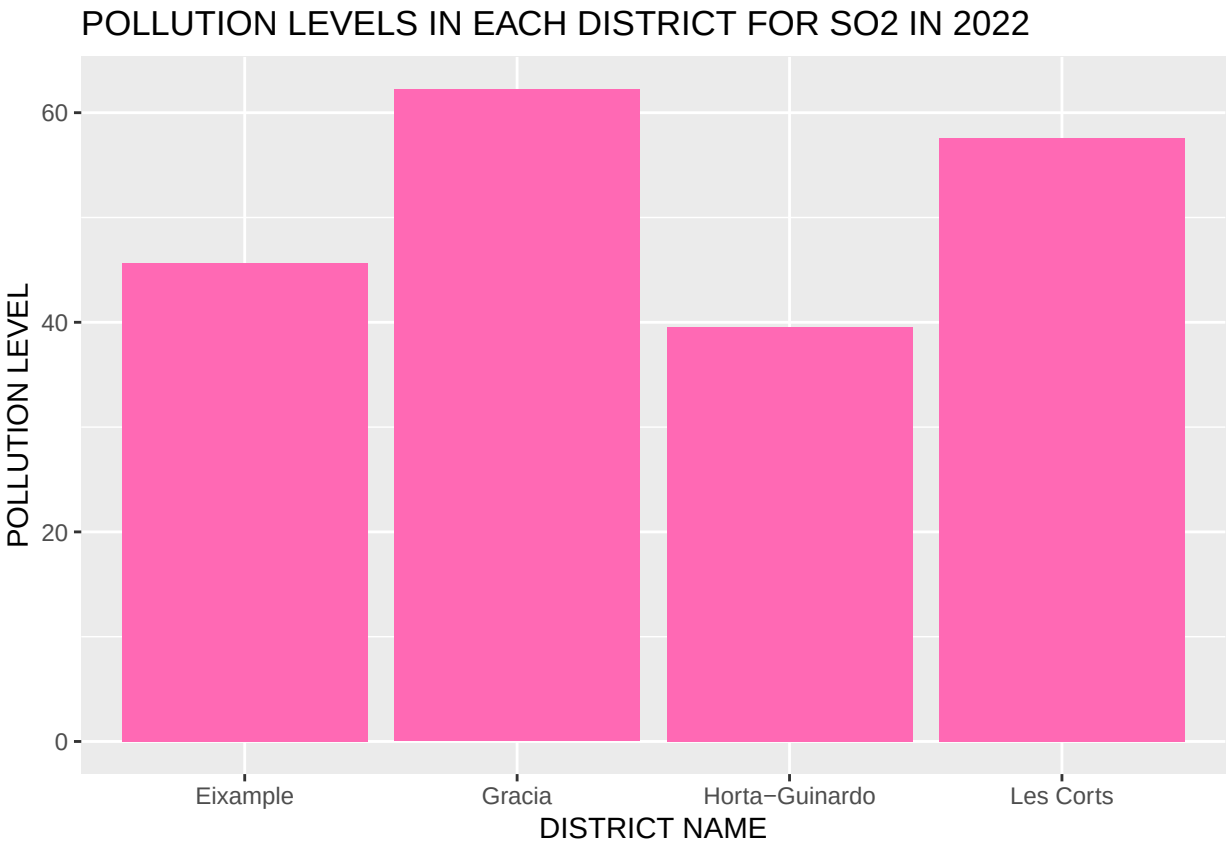
We also see peaks of NOx (a mix of nitric oxide (NO) and nitrogen dioxide (NO2)) in the area of L'Eixample being the second most present gas.

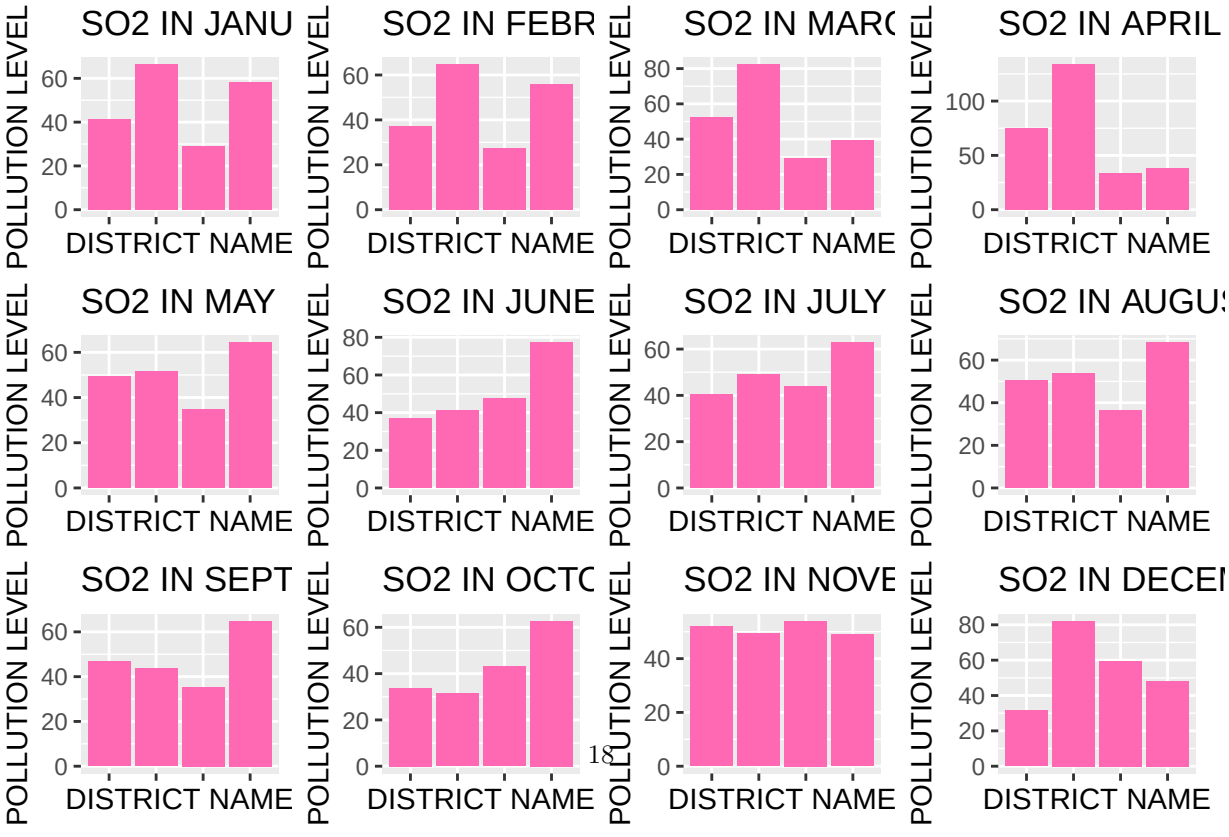
The gases less present are SO2, CO and NO, which are found only a few times during the year without great relevance in the graphs.

Throughout the months we can see that all have a large amount of polluting gases, but we could say that the month of October there is a large presence of all of them which makes it one of the most polluted months, however all are very close to par.

POLLUTION LEVELS FOR EACH CONTAMINANT DEPENDING ON THE DISTRICT

SO2





COMMENTS

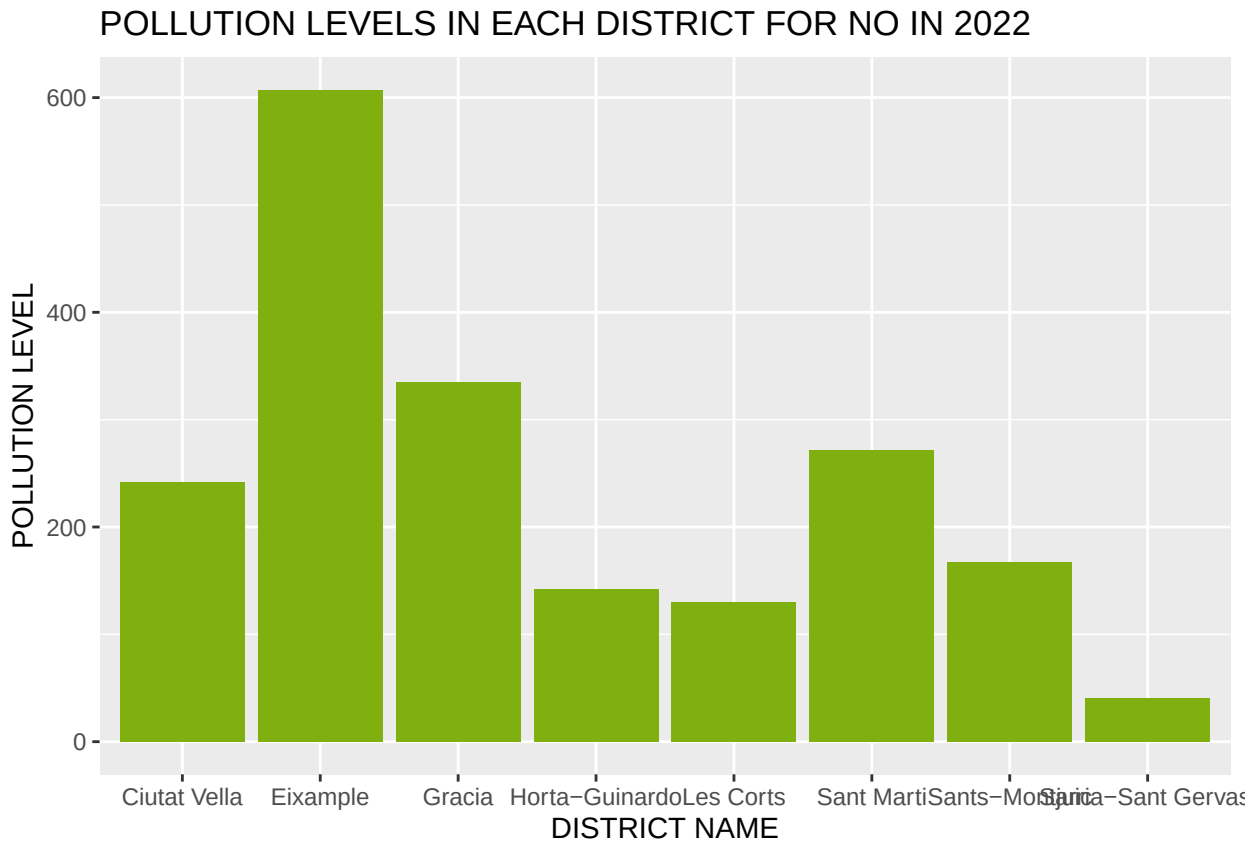
In this bar chart we can find an analysis of sulfur dioxide (SO2) in the months of the year.

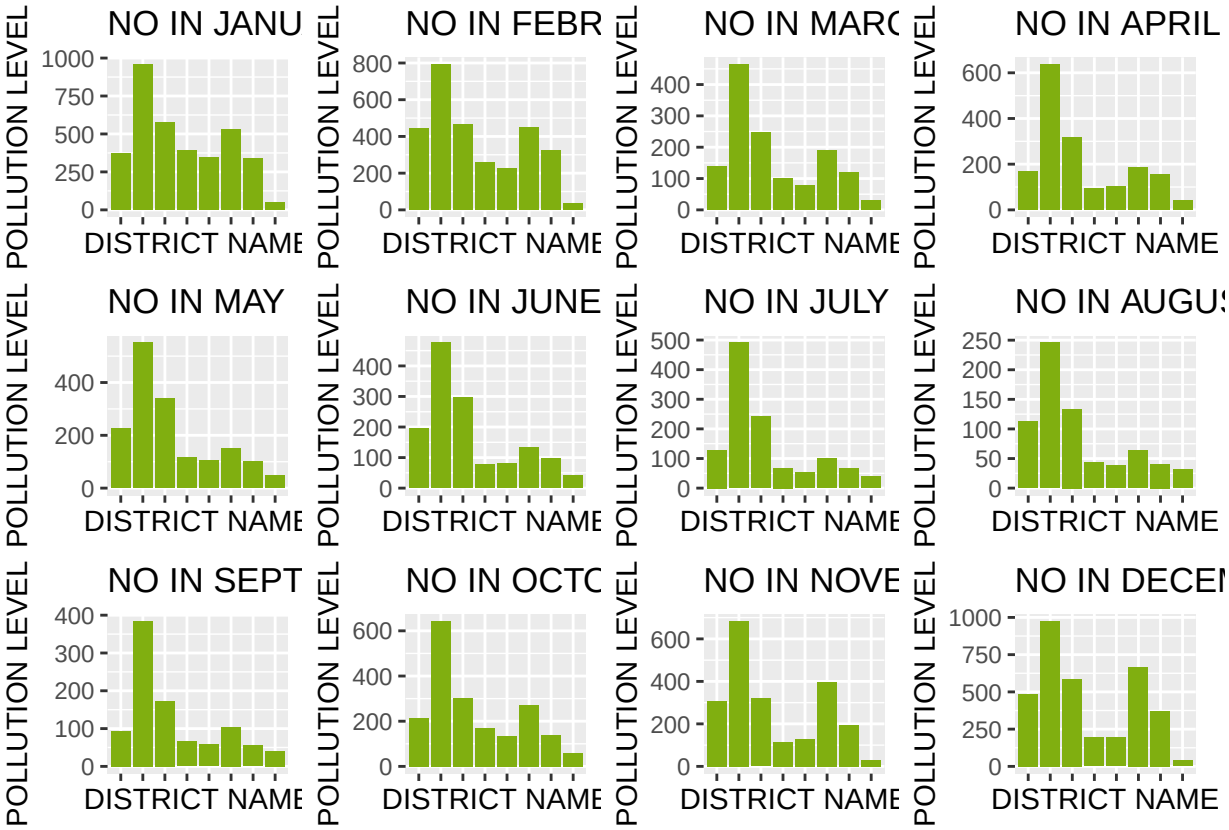
We can highlight the month of November, where in all districts the pollution has a very high level.

From July until November we can see that pollution is also very high (These months it is normal for it to be crowded since Barcelona is a tourist destination where the time of the year is not very important, normally there is always a good temperature.) and from December to April is lower.

Then we can see that Les Corts is an area that normally has a large amount of pollution throughout the year, also at Gracia district and the area of L'Eixample.

NO





COMMENTS

The bar graph does not show by months of the year the amount of nitric oxide (NO).

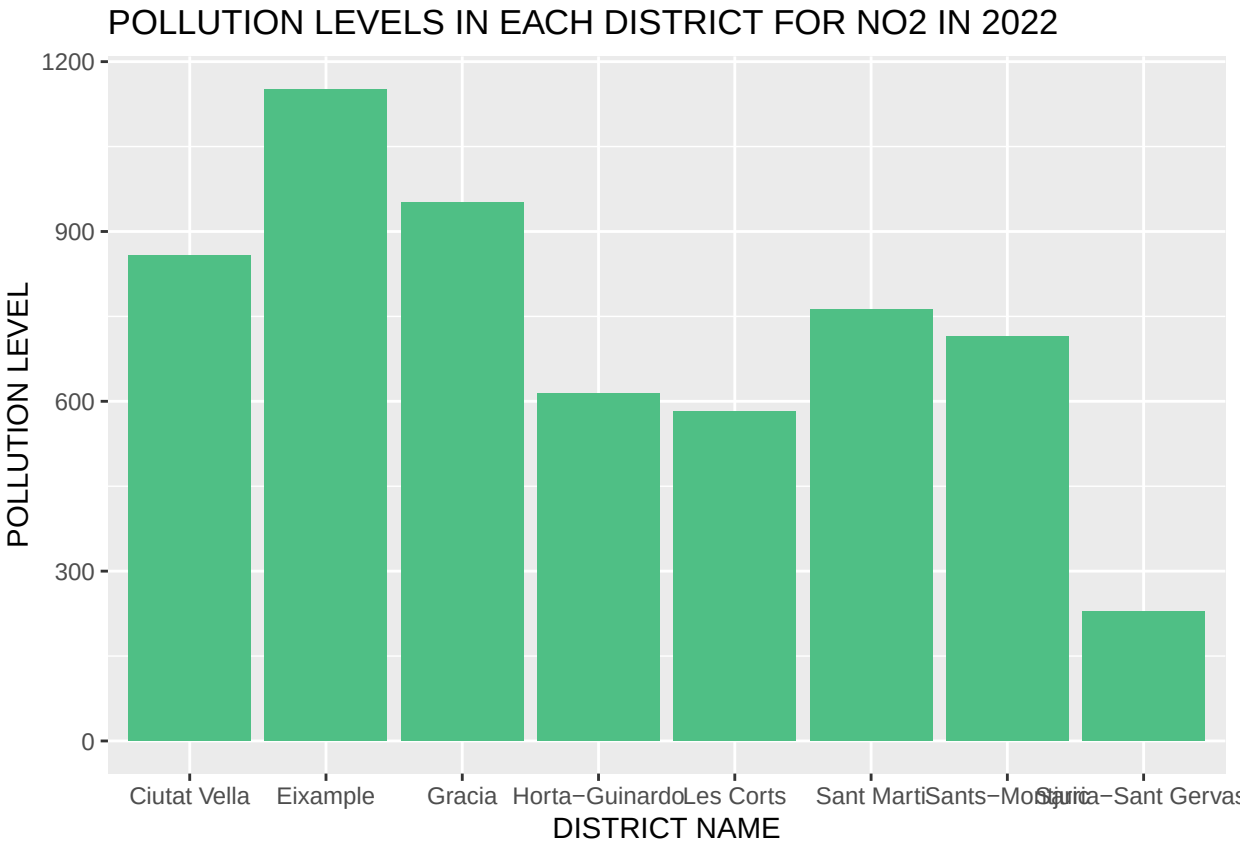
We can highlight the second bar which is l'Eixample, immediately we realize that every month of the year has very high levels of pollution, being the neighborhood with the highest exposure to this gas, none of the other areas reaches the same level as the Eixample area.

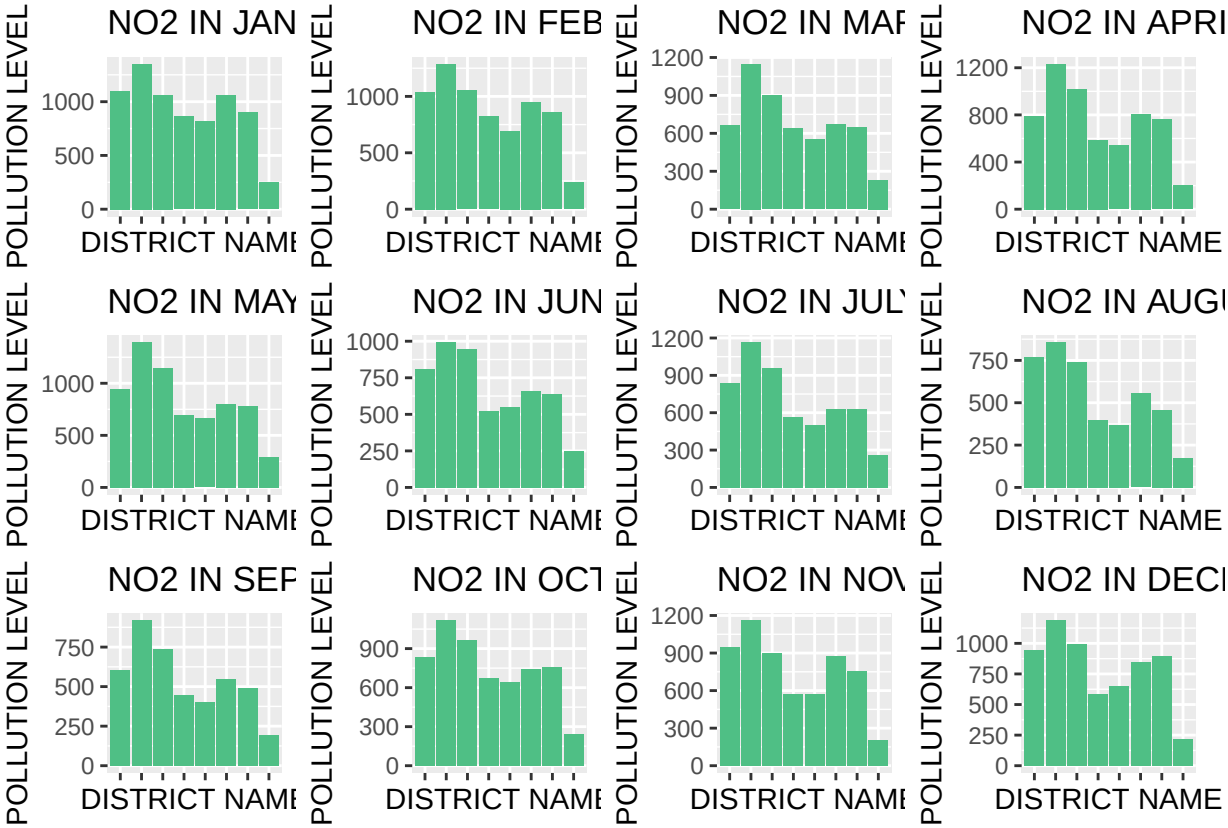
Then we can observe that the lowest months are from May to October.

In addition, the area of Sant Gervasi is always very low, which means that there is very little presence of this gas.

Finally, January is one of the most polluted months in the different zones of the country.

NO2


POLLUTION LEVELS IN EACH DISTRICT FOR NO2 IN 2022
POLLUTION LEVEL
1200
900
600
300
0
Ciutat Vella
Eixample
Gracia
Horta−Guinardo
Les Corts
Sant Marti
Sants−Montjuic
Sarria−Sant Gervas
DISTRICT NAME


NO2 IN JAN
NO2 IN FEB
NO2 IN MAR
NO2 IN APRI
NO2 IN MAY
NO2 IN JUN
NO2 IN JUL
NO2 IN AUG
NO2 IN SEP
NO2 IN OCT
NO2 IN NOV
NO2 IN DEC
POLLUTION LEVEL
DISTRICT NAME

COMMENTS

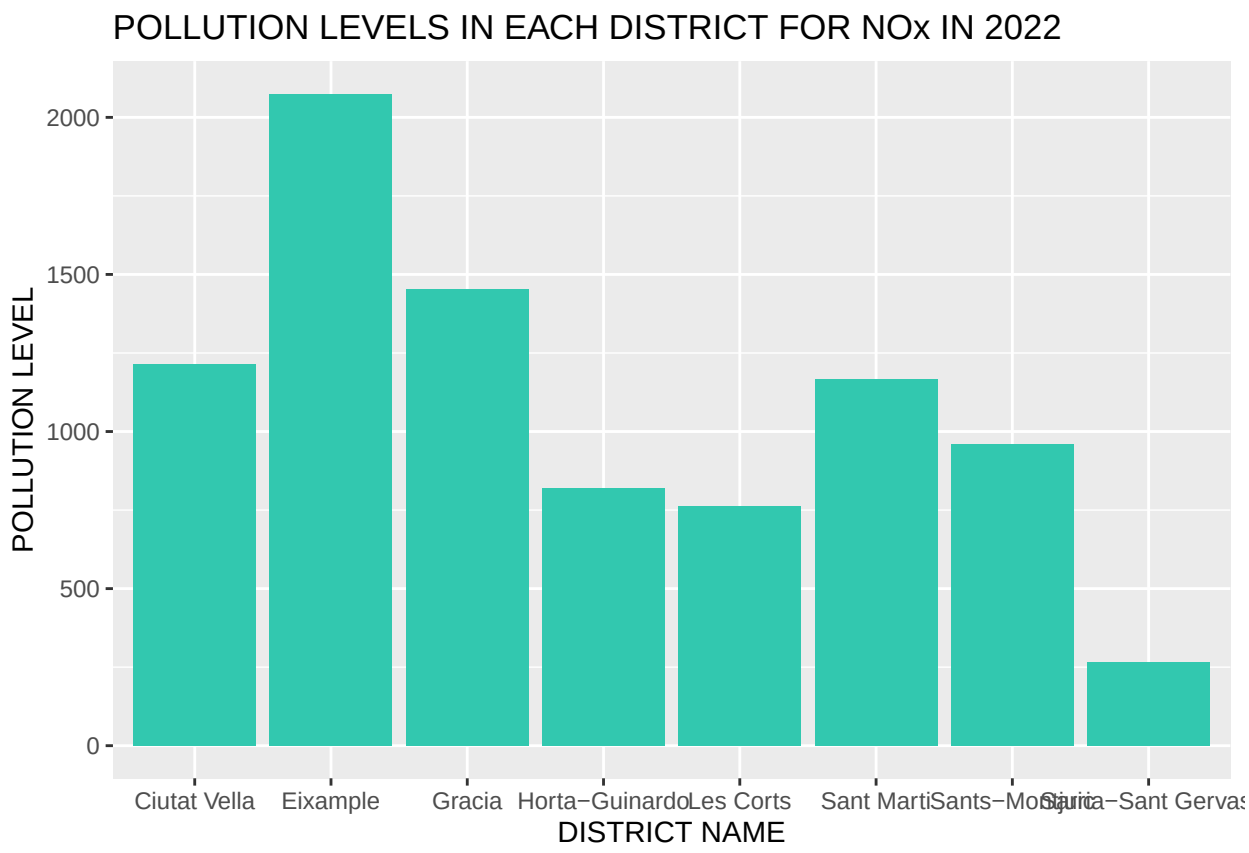
This graph shows us our the level of pollution in all months of the year emitted by nitrogen dioxide (No2).

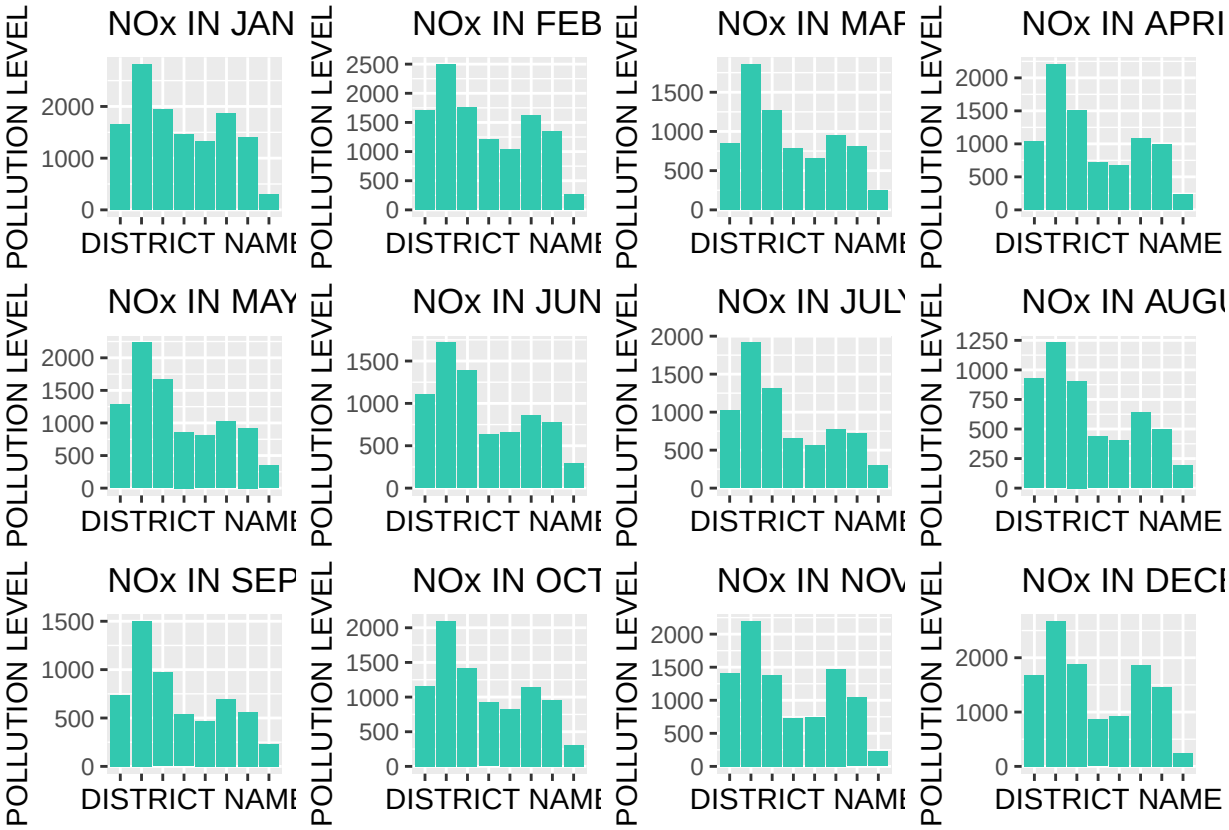
We can easily see that the whole year has a very high level of pollution, highlighting the first three which are : Ciutat Vella , L'Eixample and Gracia.

Throughout the months have always given very high levels of pollution, we can say that January is one of the most polluted and the least polluted would be a tie between September and July.

Consecutively the district of Sant Gervasi still one of the least polluted neighborhoods where we can find these gases contaminants.

NOx





COMMENTS

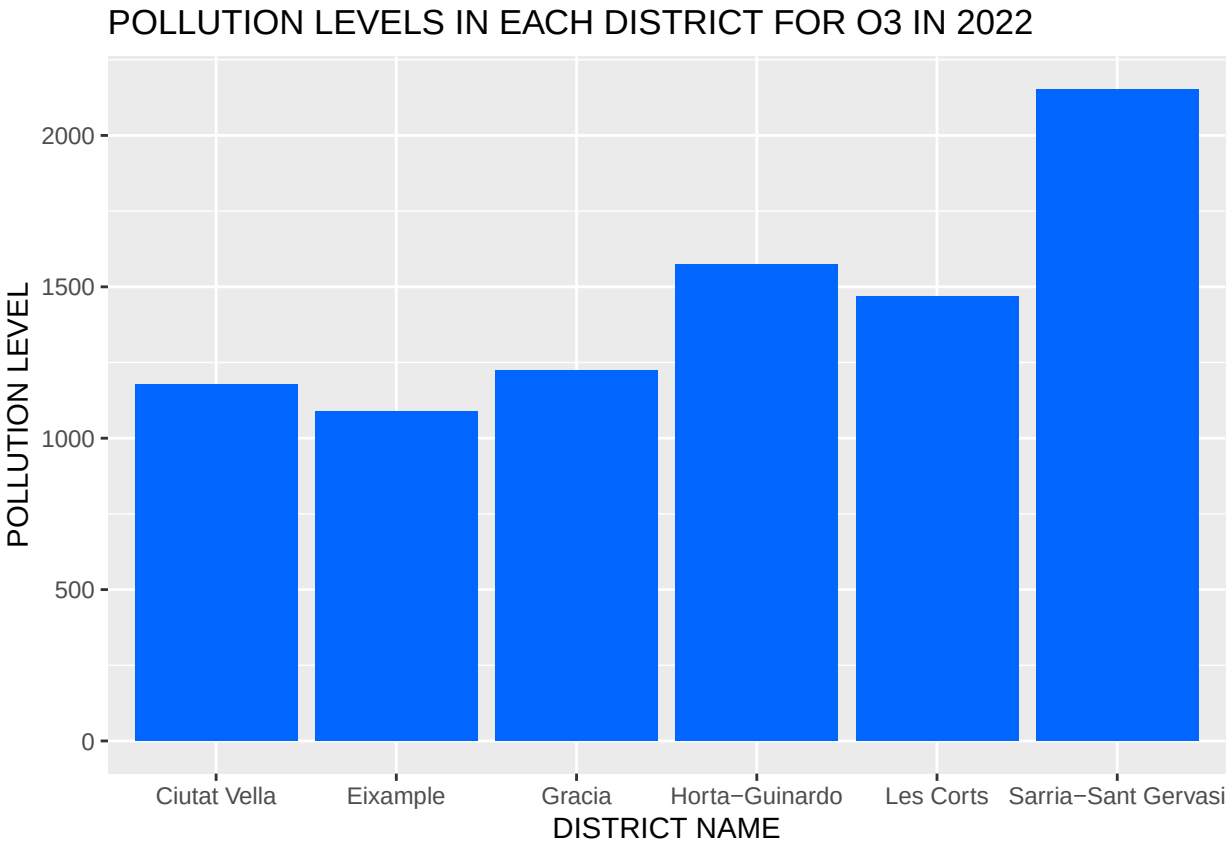
This graph analyzes the amount of NOX, ie, the set of nitric oxide (NO) and nitrogen dioxide (NO2) throughout the year 2022.

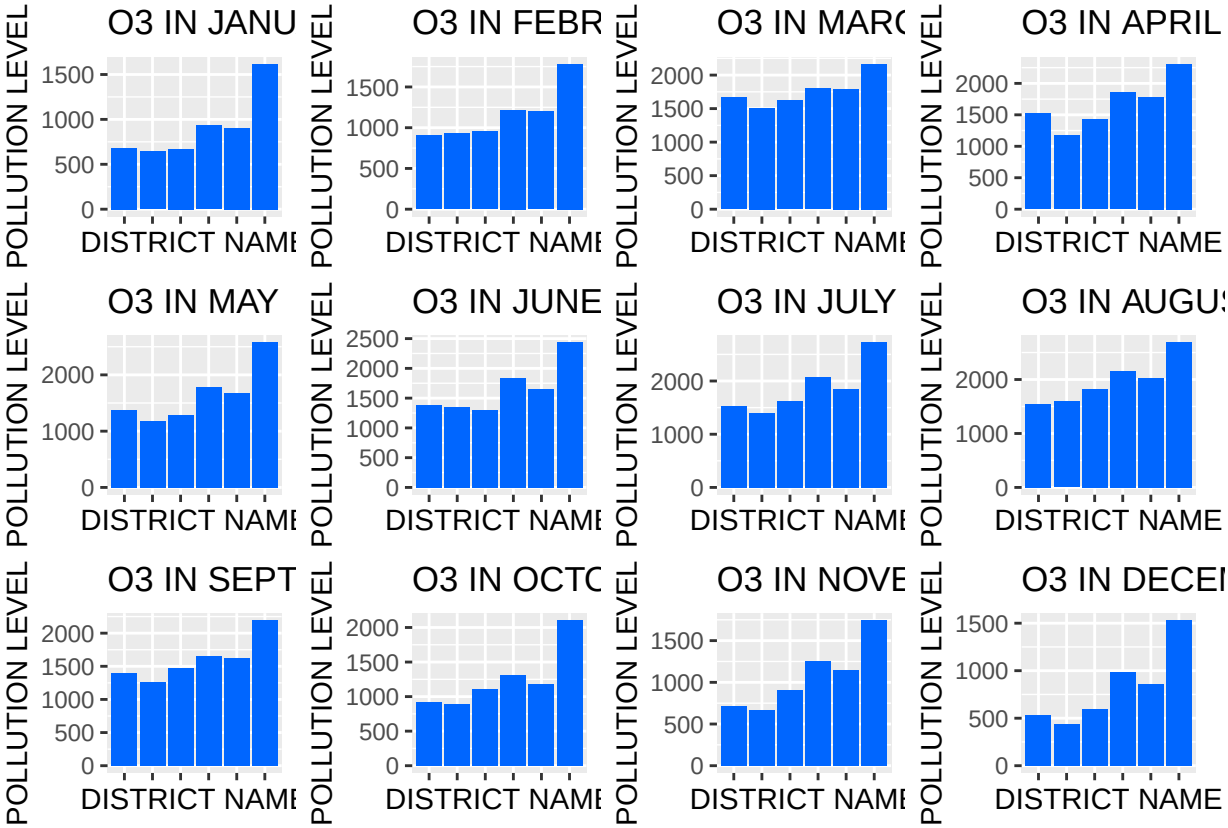
We can highlight that the Eixample neighborhood is always in the lead with the highest level of pollution in Barcelona, followed by the Sant Gervasi neighborhood, which has the lowest level throughout all months of the year.

Horta Guinardó is stable throughout all the months of the year. Both Ciutat Vella and Gracia vary throughout the year but always with high levels.

The most polluted months are the first two months of the year and the least polluted month is July.

O3


POLLUTION LEVELS IN EACH DISTRICT FOR O3 IN 2022
POLLUTION LEVEL
2000
1500
1000
500
0
Ciutat Vella
Eixample
Gracia
Horta−Guinardo
Les Corts
Sarria−Sant Gervasi
DISTRICT NAME


O3 IN JANU
O3 IN FEBR
O3 IN MARC
O3 IN APRIL
O3 IN MAY
O3 IN JUNE
O3 IN JULY
O3 IN AUGU
O3 IN SEPT
O3 IN OCTC
O3 IN NOVE
O3 IN DECE
POLLUTION LEVEL
DISTRICT NAME

COMMENTS

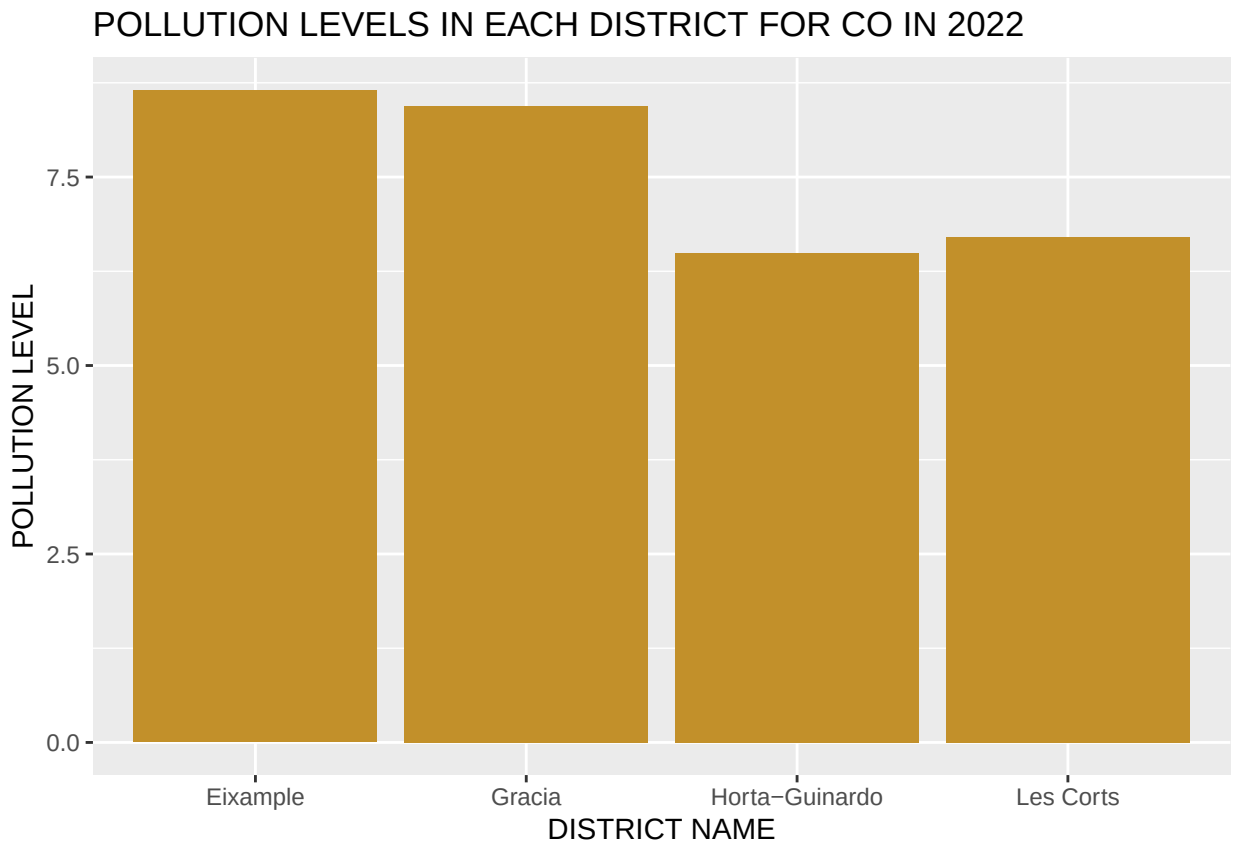
This bar graph shows us the level of O3 gas pollution i.e. ozone. We can see that it's the opposite of the other graphs we had seen.

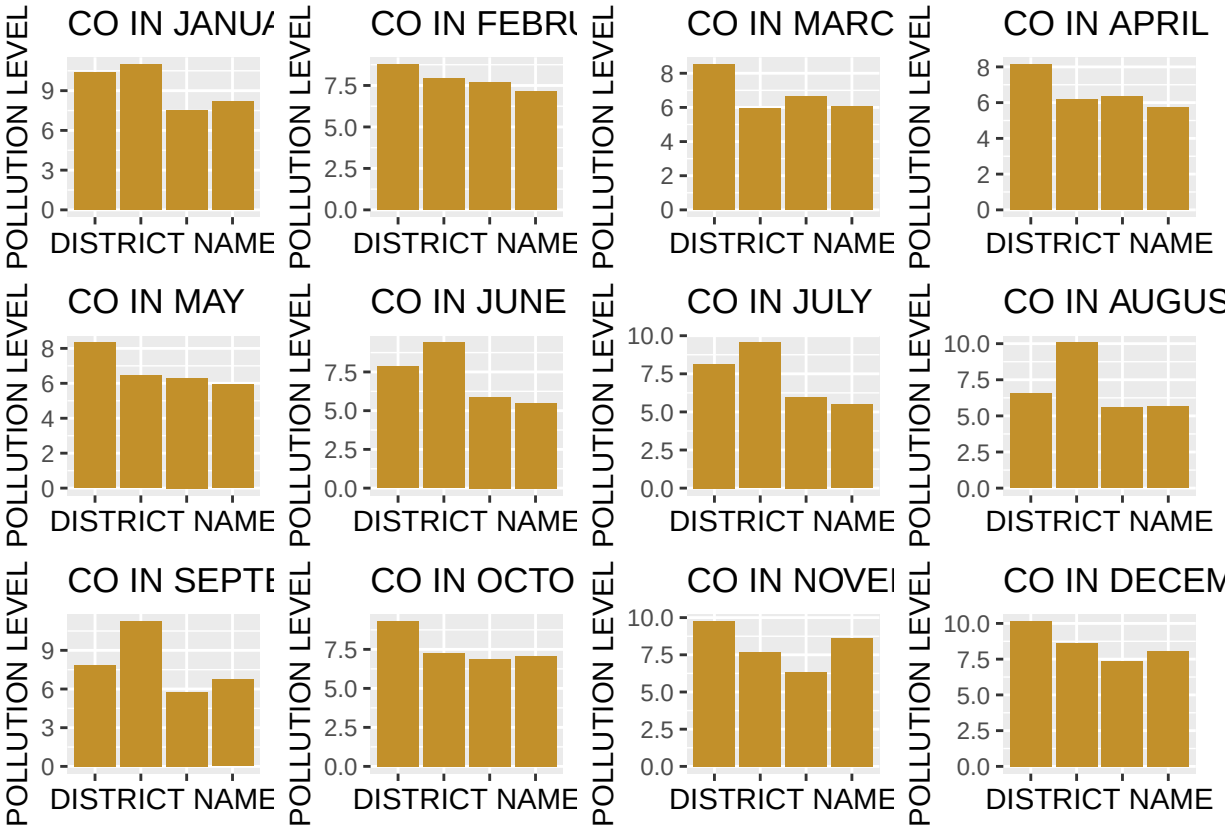
For example we can see that the neighborhood of Sant Gervasi on the previous graphs was always the minimum, now is always in the maximum exponent of pollution O3 and on the contrary we can also see that the area of l'Eixample is the area that produces the least polluting gas.

The most polluted month we could say that is the month of March and the least would be the month of December and January.

Finally to emphasize that the district of Horta Guinardó (in fourth position) also has very high levels throughout the year.

CO





COMMENTS

In this graph we can find the amount of carbon monoxide (CO) in different districts of Barcelona throughout the year 2022.

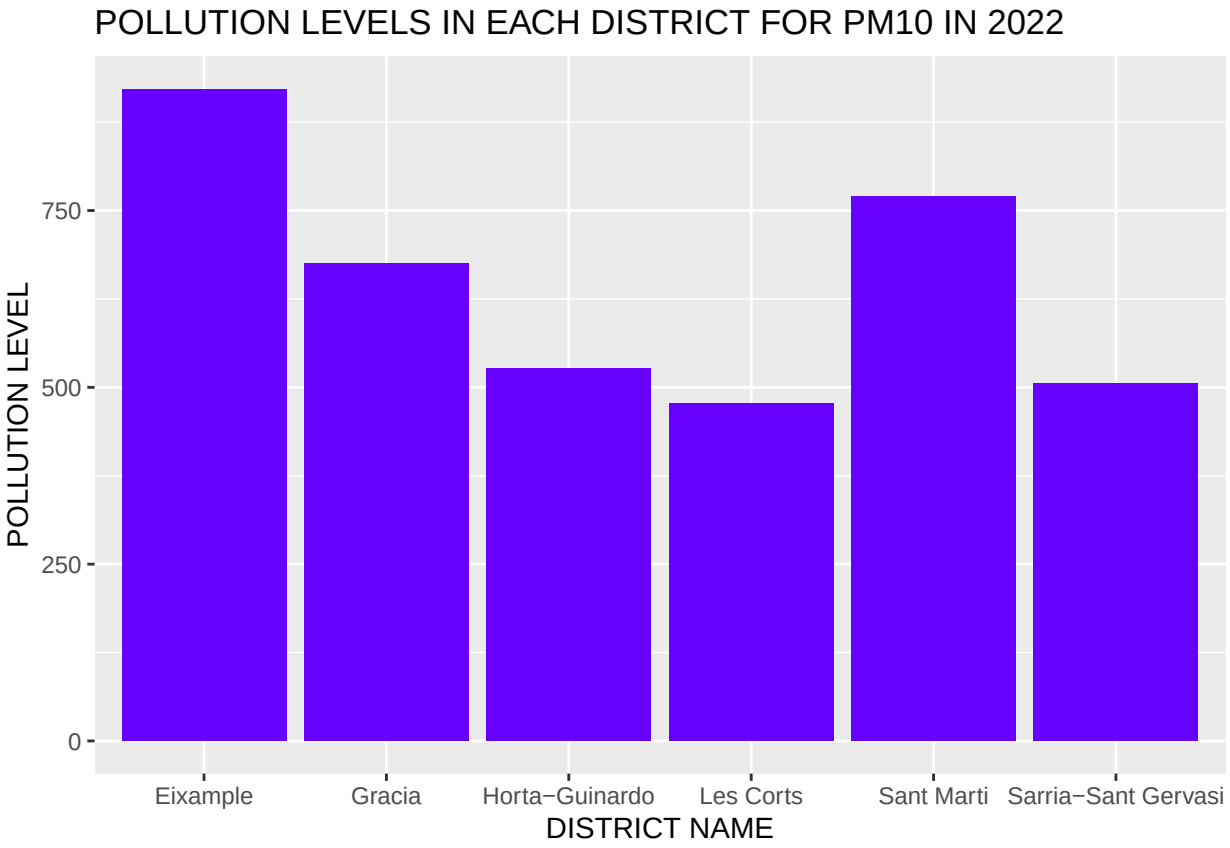
At first glance in the first graph, we can see how the four neighborhoods abound this gas, especially in the area of L'Eixample and Gracia. We can see that pollution levels are very high throughout the year.

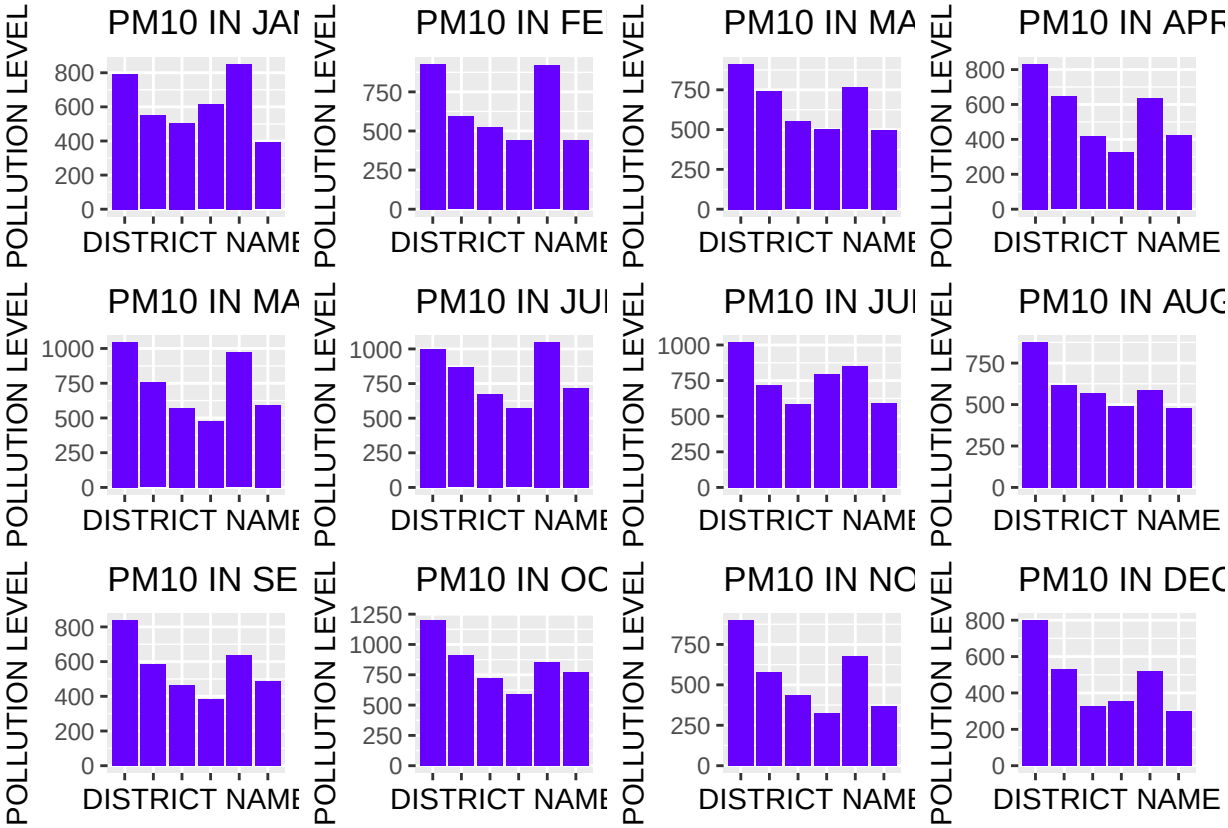
The month with the highest presence of this gas is February, where in the 4 neighborhoods it has very high levels of CO.

The Eixample area has a total of 7 peaks, being more than half of the year emitting large amounts of pollution. The area of Gracia is not far behind either and has a total of 5 peaks, however the levels are never low.

The area of Les Corts and Horta are also very close, but also have very high levels of pollution, the lowest level is found in the area of Horta with a 6 (of pollution level).

PM10


POLLUTION LEVELS IN EACH DISTRICT FOR PM10 IN 2022
POLLUTION LEVEL
0
250
500
750
Eixample
Gracia
Horta−Guinardo
Les Corts
Sant Marti
Sarria−Sant Gervasi
DISTRICT NAME


PM10 IN JAI
PM10 IN FE
PM10 IN MA
PM10 IN APF
PM10 IN MA
PM10 IN JUI
PM10 IN JUI
PM10 IN AUG
PM10 IN SE
PM10 IN OC
PM10 IN NO
PM10 IN DEC
POLLUTION LEVEL
DISTRICT NAME

COMMENTS

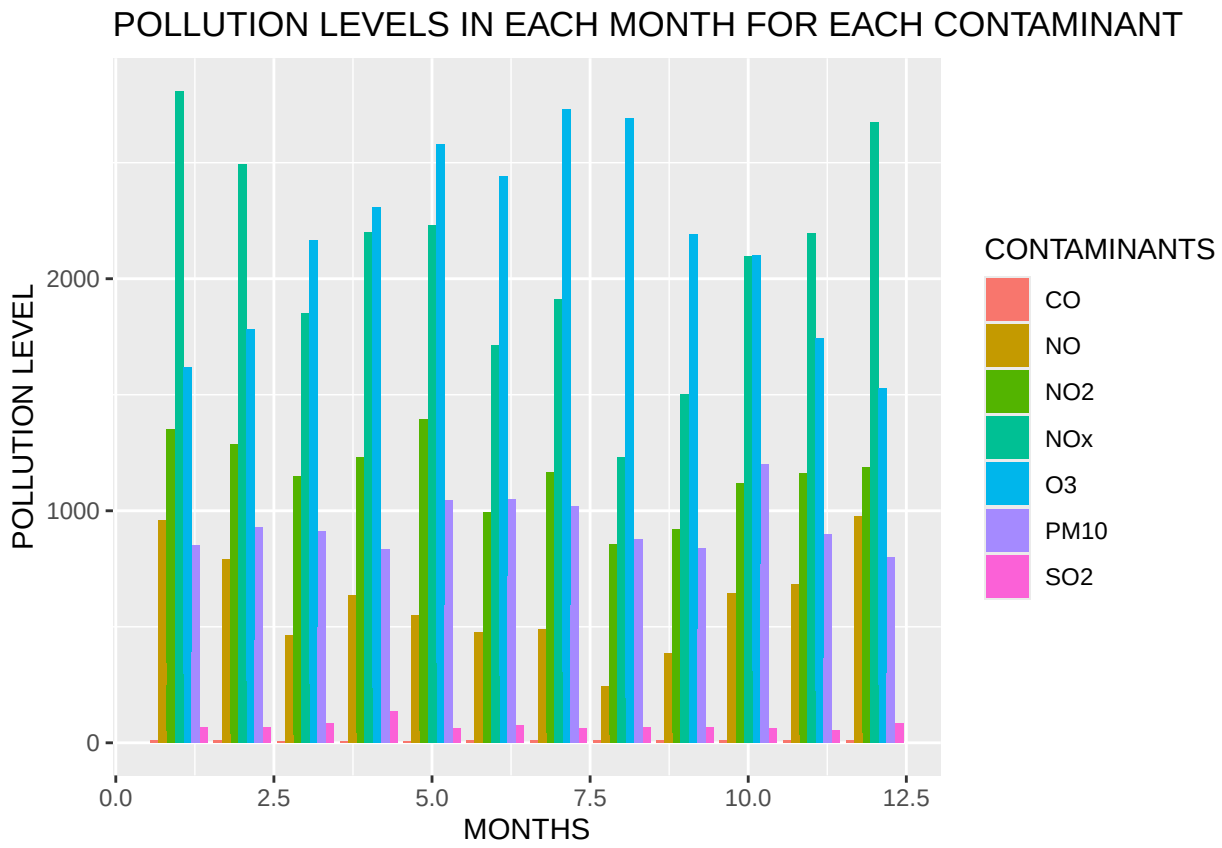
The graph shows the amount of PM10 (PM10 is a small solid or liquid particles of dust, ashes, metallic particles, cement, pollen, among others) found in different districts of Barcelona.

We can see that throughout Barcelona there is a large amount of these particles especially in L' Eixample and Sant Martí areas. The areas that contain less are: Les Corts, Sant Gervasi and Horta.

Throughout all the year we can see that November and December are the months with the lowest levels and the highest from January to the end of July.

L' Eixample area has peaks every month of the year except January. Also the area of Sant Martí has four peaks throughout the year, putting it in second position of the most polluted districts.

FOR EACH MONTH, WE PLOT ALL THE POLLUTION LEVELS FOR ALL CONTAMINANTS



COMMENTS

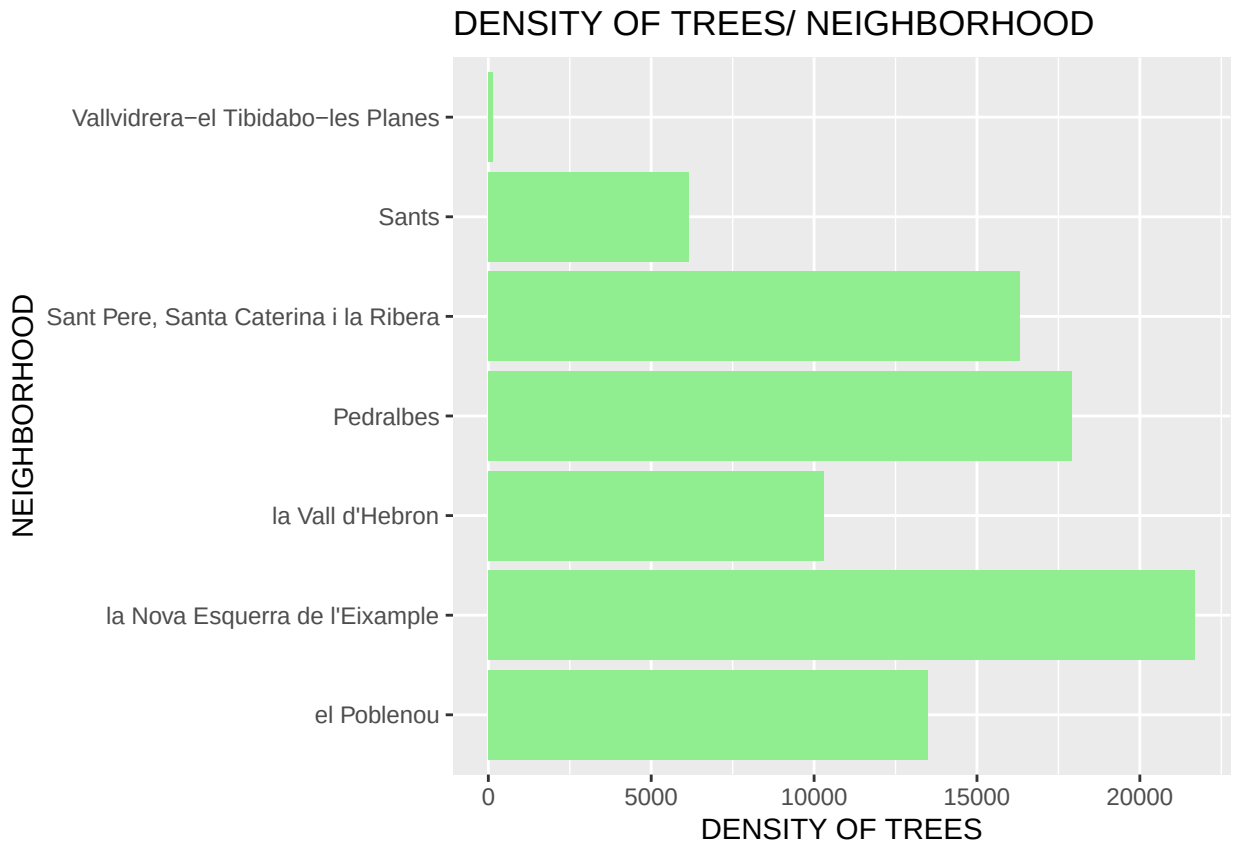
This plot is a bar chart. The data used in the plot comes from a series of dataframes, each one representing the pollution level in a specific month. The dataframes are named "stations_pollution_JANUARY", "stations_pollution_FEBRUARY", "stations_pollution_MARCH" and so on.

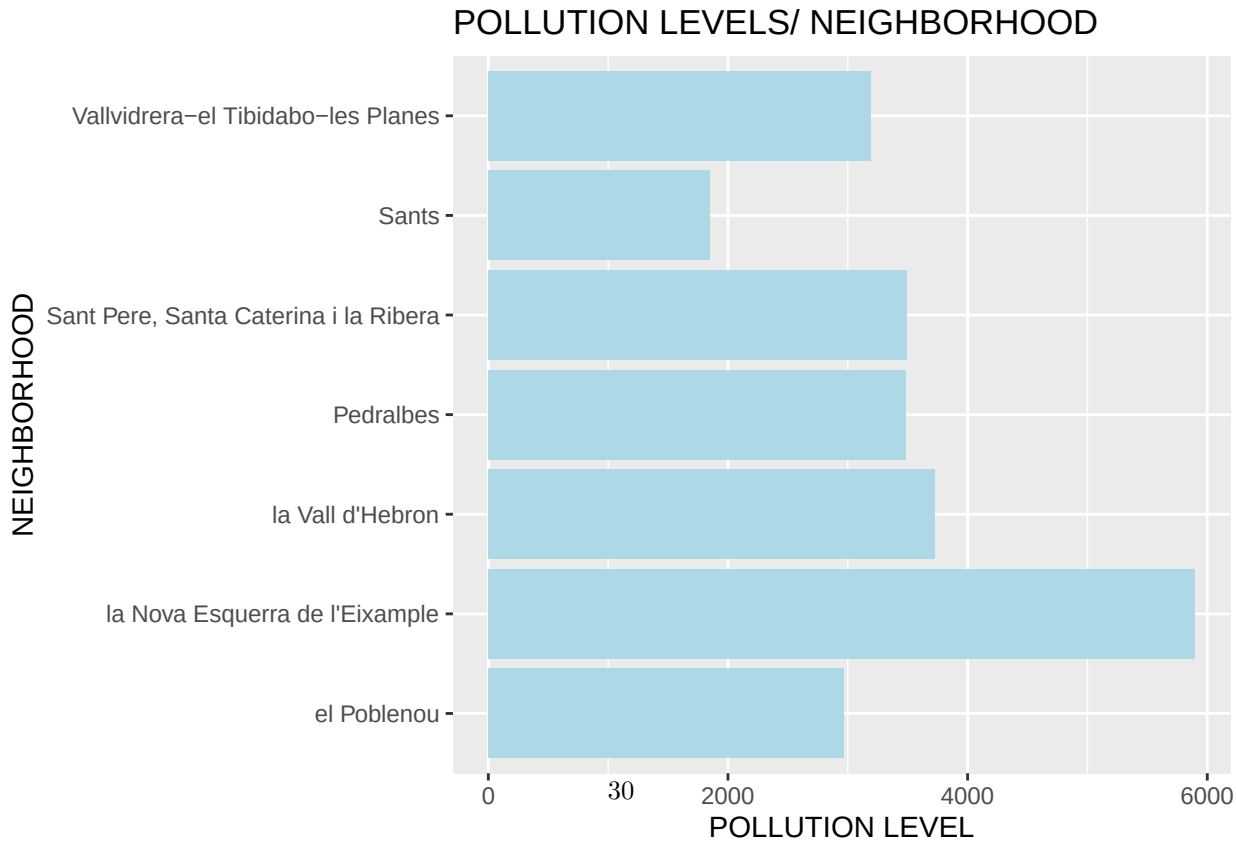
First, it adds a new column called "Month" to each dataframe, with a value representing the corresponding month (1 for January, 2 for February, 3 for March, and so on). Then, it uses the union function to combine all these dataframes into one dataframe called "pollution_by_months".

The plot shows the pollution levels for each month and for each contaminant. The x-axis represents the months, the y-axis represents the pollution level and the fill of the bars represents the type of contaminant. It uses geom_bar and stat="identity" to create the bars. The position="dodge" makes sure that each bar is plotted separately. The plot also includes labeled axes for Months, Pollution Level, and Contaminants and a title "POLLUTION LEVELS IN EACH MONTH FOR EACH CONTAMINANT". This plot can be used to observe the trend of pollution levels for each month and for each contaminant.

SECOND PART OF THE PROJECT

DENSITY OF TREES AND POLLUTION LEVEL BY NEIGHBORHOODS





COMMENTS

The first plot is a bar chart that represents the density of trees in each neighborhood. The second plot is also a bar chart that represents the pollution level in each neighborhood.

Both plots use the same dataframe and the same x-axis, which represents the neighborhood name, and both plots are ordered by this variable. The first plot represents the density of trees with a y-axis and it is filled with light green color. The second plot represents the pollution level with a y-axis and it is filled with light blue color.

We have put them together as it is important to see the difference between one and the other.

A very important thing to highlight is the area of Tibidabo we can see that it has hardly any trees, therefore makes its pollution level is very high, for example the area of Sants has a high level of pollution however to have more density of trees makes the pollution reduced.

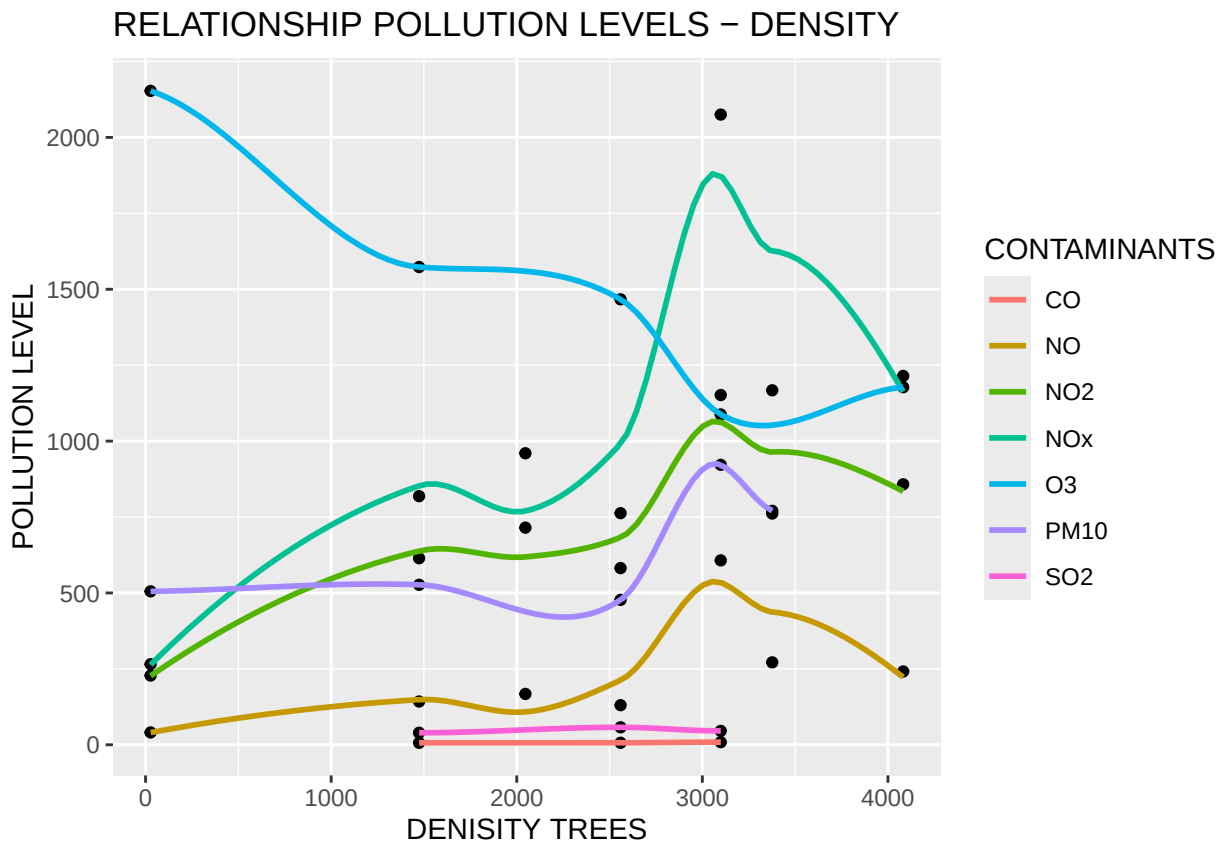
Also to highlight, we have the area of l'Eixample that although there is a high density of trees the level of pollution is even higher so the right thing to reduce the level of pollution would be to put more trees in that area.

Then the area of Pedralbes and Sant Pere, which are at the same level in pollution, we can see that its density of trees is much higher than the level of pollution.

Finally, the area of Vall d'Hebron has a lower density of trees than the pollution density, however, as we have already mentioned, as it is an area with a lot of car traffic, which is an important factor to take into account.

In conclusion to this graph, we can deduce that the greater the amount of trees per neighborhood, the lower the level of pollution will be, although in areas such as l' Eixample it has a higher level, however if we put a greater amount of trees we could solve this problem.

RELATIONSHIP BETWEEN POLLUTION LEVELS AND DENSTITY TRESS



COMMENTS

This is a scatter plot. It takes data from a dataframe called "backstreet_trees" and creates a scatter plot of the "Density_trees" column on the x-axis and the "POLLUTION_LEVEL" column on the y-axis. The plot also adds a smooth line through the points, which is colored by the "Desc_Contaminant" column. The plot also includes labeled axes for Density Trees and Pollution Level and a title "RELATIONSHIP POLLUTION LEVELS - DENSITY". This plot is used to investigate the relationship between density of trees and pollution levels in different neighborhoods, and how the pollution levels vary depending on the type of contaminant.

Firstly, the higher the density of trees, the lower the number of O3 gas contamination will be. However the NOx we can see that the higher the density of trees the higher the amount of pollution level, however from 3000 it begins to decline.

Then the NO2 and PM10 are quite similar in the path they make, however at the beginning the PM10 remain neutral until the anxiety of trees about 2500 which is when it begins to grow making a peak of about 900 approx, and then fade. On the other hand the NO2 has a maximum peak of more than 1000, it is quite similar to the rest, at the beginning from 2500 it starts to grow and then it decays again.

Finally, SO2 and CO remain neutral at the beginning and from 1500 it grows a little but then declines again and disappears.

RELATIONSHIP BETWEEN DENSITY TREES, TOTAL POLLUTION AND NEIGHBORHOOD

Click here to see the 3D Plot.

COMMENTS

We have created a 3D scatter plot using the plotly library. It takes data from a dataframe called "backstreet_trees" and performs several operations on it. First, it creates a new column called "Density_people" by dividing the "Population" column by the "Size" column. Then it groups the data by neighborhood name and calculates the mean density of trees and the total pollution for each neighborhood.

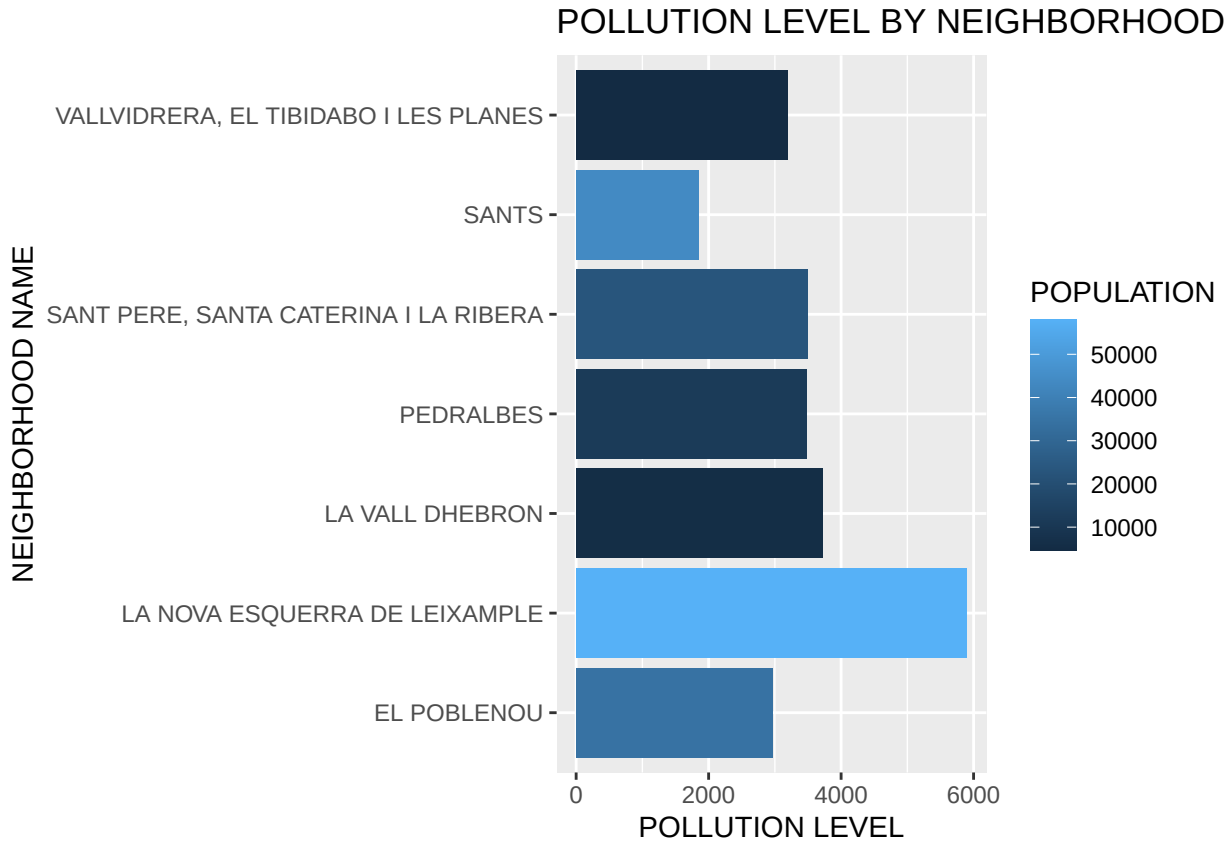
The resulting data is then used to create a 3D scatter plot with the x-axis representing the mean density of trees, y-axis representing the total pollution and z-axis representing the neighborhood. The plot is also colored by the neighborhood name. The layout of the plot includes labeled axes for Density Trees, Pollution and Neighborhoods.

If we start analyzing the 3D plot, the highest point it's in the neighborhood of la Nova Esquerra de l'Eixample having over 5800 of pollution level and over 3000 of density of trees. It has a lot of density of trees as it has 3 public parcs and gardens and arround 8 mini-gardens (interiors d'illa).

As you can see there's no direct relation between the density of the trees and the pollution level as some other factors should be included in the analysis such as the density of the people in this neighborhood which will be later analyzed.

THIRD PART OF THE PROJECT: PEOPLE DATASETS

POLLUTION LEVEL BY NEIGHBORHOOD



COMMENTS

This plot displays the average pollution level in different neighborhoods, represented by the y-axis. The bars are grouped by the neighborhood name on the x-axis and the fill color of the bars represent the population of the neighborhood. The graph is flipped horizontally to better show the neighborhood names. The title of the graph clearly states the information being displayed. Overall, the plot effectively presents the data in a clear and easy-to-understand way.

To highlight we could put La Nova Esquerra de l'Eixample as we have commented previously is a very busy area as it is in the center of Barcelona, has a total of 6000.

Then we have the area of Vall d'Hebron, being a place where hundreds of cars pass daily is normal that the level of pollution is so high with almost 4000.

In addition, we can see that Pedralbes and Sant Pere are almost tied. Pedralbes is located in the district of Les corts and Sant Pere is located in Ciutat Vella. As we have already mentioned the area of Les Corts is very crowded due to the infrastructures that are found there, however the area of Ciutat Vella is one of the most crowded of Barcelona, we can find La Rambla, the Gothic Quarter, the Olympic Port, among others. Being an area so well communicated (referring to public transport), therefore it may be the reason why the level of pollution is the same as Les Corts, an area not so well communicated.

We can also see Tibidabo area (+3000 pl) and Poblenou at San Marti district (3000 pl) have a very similar level of pollution. However, the San Martí area is very well communicated, which may justify the difference

between the two districts.

Finally we have the area of Sants, quite busy as it is in the center of Barcelona, being an area very well connected with public transport as AVE, train, subway or bus, it can be one of the reasons why it has a lower level of pollution.

RELATIONSHIP BETWEEN DENSITY PEOPLE, POLLUTION LEVEL AND NEIGHBORHOOD

Click here to see the 3D Plot.

COMMENTS

This plot is a 3D scatter plot that shows the relationship between pollution level, density of people, and neighborhood. The x-axis represents the pollution level, the y-axis represents the density of people, and the z-axis represents the name of the neighborhood. Each point on the plot represents a neighborhood, with the color of the point indicating the neighborhood name. The plot uses the library plotly to create the 3D scatter plot, and the layout of the plot includes a title and labels for each axis. The plot is showing the relationship between pollution level, density people and neighborhood in an interactive way, enabling the user to rotate and zoom in and out of the plot to get a better understanding of the data.

Moving to the analysis of the plot, the highest point it's reached in the neighborhood of la Nova Esquerra de l'Eixample setting the pollution level over 5000 and the mean density of people arround 43000. Comparing this data with the other maps and data from the Ajuntament of Barcelona, we see that in this neighborhood we have lots of educational centers what means a lot of movement in terms of transport (public or private). Also there's 4 residences for elderly people and around 30 cultural and leisure facilities.

In the neighborhood Sants we have a high mean density of people, a low pollution level and a high density of trees. This can be because we have 8 gardens and parcs around the neighborhood. Also it do not have a lot of toursim as it is on the left part of Barcelona and the only point that attracts people there is the Station of Sants, the principal train station in Barcelona. Nowadays, as pollution is one of the biggest concerns of the city of Barcelona and of the Catalan and Spanish government, most of the public transports are non-pollutant or less pollutant than normal vehicles so, here we have the explanation for the low pollution level.

We can also observe from this 3D plot that there's a direct relationship between the pollution level and the density of people as we can observe that in the neighborhoods less populated the level of pollution drops.

```r
sessionInfo()
```

```
## R version 4.3.3 (2024-02-29 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 11 x64 (build 22631)
##
## Matrix products: default
##
##
## locale:
## [1] LC_COLLATE=Spanish_Spain.utf8  LC_CTYPE=Spanish_Spain.utf8
## [3] LC_MONETARY=Spanish_Spain.utf8 LC_NUMERIC=C
## [5] LC_TIME=Spanish_Spain.utf8
##
## time zone: Europe/Madrid
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] patchwork_1.2.0 lubridate_1.9.3 forcats_1.0.0   stringr_1.5.1
##  [5] dplyr_1.1.4     purrr_1.0.2     readr_2.1.5     tidyr_1.3.1
##  [9] tibble_3.2.1    ggplot2_3.5.1   tidyverse_2.0.0
##
## loaded via a namespace (and not attached):
##  [1] utf8_1.2.4        generics_0.1.3    lattice_0.22-5    stringi_1.8.4
##  [5] hms_1.1.3         digest_0.6.35     magrittr_2.0.3    evaluate_0.23
##  [9] grid_4.3.3        timechange_0.3.0  fastmap_1.2.0     Matrix_1.6-5
## [13] mgcv_1.9-1        fansi_1.0.6       scales_1.3.0      cli_3.6.2
## [17] rlang_1.1.3       crayon_1.5.2      splines_4.3.3     bit64_4.0.5
## [21] munsell_0.5.1     withr_3.0.0       yaml_2.3.8        tools_4.3.3
## [25] parallel_4.3.3    tzdb_0.4.0        colorspace_2.1-0  vctrs_0.6.5
## [29] R6_2.5.1          lifecycle_1.0.4   bit_4.0.5         vroom_1.6.5
## [33] pkgconfig_2.0.3   pillar_1.9.0      gtable_0.3.5      glue_1.7.0
## [37] xfun_0.44         tidyselect_1.2.1  highr_0.10        rstudioapi_0.16.0
## [41] knitr_1.46        farver_2.1.2      nlme_3.1-164      htmltools_0.5.8.1
## [45] rmarkdown_2.27    labeling_0.4.3    compiler_4.3.3
```