VDS2425 Project Report

Part 1. Metadata

- Students: student 1 (Sadia Parveen, 2469624), student 2 (Muhammad Umair, 2469710), student 3 (Naheed Hidayat, 2469623, university), student 4(Laiba Tahir, 2469634)
- Dataset: Madrid
- Group number: 12

Part 2. Project description

Project/Data Description:

This project investigates air pollution in Madrid through exploratory data analysis, aiming to raise awareness about its patterns and trends. We analyze multiple pollutant factors to understand how pollution has evolved over time, identify the areas with the highest and lowest levels, and track whether individual pollutants have improved or worsened. The goal is not just to answer these questions, but to explore the data visually and uncover insights about the city's pollution status.

Features Description:

The key features include the station ID, station name, longitude and latitude, date, and various pollutant types. The station ID is the key that connects two different files: one contains information about each station (like location), and the other has pollution readings. The station name makes it easier to identify locations instead of remembering long ID numbers. The date is also a key feature — it helps us group and analyse pollution levels by day, month, or year.

Questions that will guide us in the rest of the project.

1) How has pollution in Madrid evolved over time (2001–2018)?
2) Which areas of the city are most and least polluted, and how has this changed?
3) Are there seasonal or recurring patterns in pollutant levels across years?

Part 3. Design

To answer the four main questions in the project, we first tried to understand what exactly each question was asking and what kind of patterns or comparisons would help answer it. Instead of jumping straight to the final charts, we explored the data step by step, and let the design evolve based on what we found.

3.1 Operationalisation: from questions to tasks

Question 1:

- **Title**: Pollution in Madrid Over the Years
- **Question**: How has pollution in Madrid evolved between 2001 and 2018?
- **Action**:
    - The data is spread across multiple CSV files (one for each year). First, we merge all files into a single dataset and add a new column year extracted from the date column.
    - Since each pollutant is measured on a different scale, we normalize the data using **Z-score normalization**. This allows us to compare trends across different pollutants.
    - After cleaning and normalizing, we compute yearly and monthly averages for key pollutants to explore how their levels have changed over time.
- **Target**: The evolution of **pollutant levels** across years (2001–2018).
- **Objects**:
    - Pollutant values (PM10, NO_2, O_3, CO, SO_2, PM2.5, etc.)
    - Station data (from multiple files)
    - Date information (used to extract month and year)
- Measure
    - Normalized concentrations of pollutants
    - Yearly and monthly averages
- Groupings:
    - By year and month for time-based trends
    - By pollutant type to compare how each pollutant behaved over time

Question 2:

- Title: Most and Least Polluted Areas in 2018
- Question: Which are the areas of Madrid where pollution is highest / lowest in 2018?
- Action:
    - Filter the full dataset to keep only the data from **2018**.
    - Calculate the **average value of each pollutant per station**.
    - Data is normalized already in previous question.
    - Now we can perform two things:
        - Calculate a **composite pollution score** per station using the average of normalized pollutant values, and
        - Visualize multiple pollutants separately per station.
    - Rank the stations based on pollution score and compare visually (e.g., bar chart or map view).
- Target: Identifying **stations (areas)** with highest and lowest pollution levels in 2018.
- Objects:
    - Station-wise pollutant readings from 2018
    - Station metadata (name, coordinates)
- Measure:
    - Average pollutant concentrations per station
    - Normalized values/ composite pollution index

- Groupings:
    - Grouped by **station ID / name**
    - Optionally by **pollutant type**

Question 3:

- Title: Changes in Pollution by Area (2008 vs 2018)
- Question: Which are the areas of Madrid where pollution has improved / worsened more between 2008 and 2018?
- Action:
    - Filter the dataset to keep only **2008** and **2018**.
    - Calculate **station-wise average** pollutant levels for both years.
    - Already normalized data
    - For each station, calculate the **difference** between 2008 and 2018 average values for each pollutant.
    - Visualize the differences per station — which areas improved (lower pollution) or worsened (higher pollution).
- Target: Measuring **pollution change over time per area**.
- Objects:
    - Station readings for 2008 and 2018
    - Each pollutant per station
- Measure:
    - Change in pollutant level per station (Δ = 2018 - 2008)
- Groupings: Characteristics that separate the data
    - Grouped by **station**
    - Grouped by **pollutant type**

Question 4:

- Title: Evolution of Each Pollutant Type
- Question: How have the different measurements of pollution evolved between 2008 and 2018?
- Action:
    - Filter the dataset to **2008–2018**.
    - For each **pollutant**, calculate monthly and yearly averages across the entire city.
    - Use visualizations (e.g., line plots, heatmaps) to show seasonal trends and long-term changes.
    - Compare pollutants side-by-side to see which ones improved, stayed stable, or worsened.
- Target: Understanding **how each pollutant has changed over time**.
- Objects:
    - Pollutant measurements across years and months
    - Entire city (not per station)
- Measure:
    - Monthly and yearly averages per pollutant
- Groupings:

- Grouped by **pollutant type**
- Grouped by **year** and **month**

3.2 The design process

To start, we sketched around 12 different ideas to explore pollution in Madrid — things like trends over time, differences between areas, and seasonal changes. These quick sketches helped us think through different ways to visualize the data. Later, we picked a few that stood out and added notes to show how they could work better.

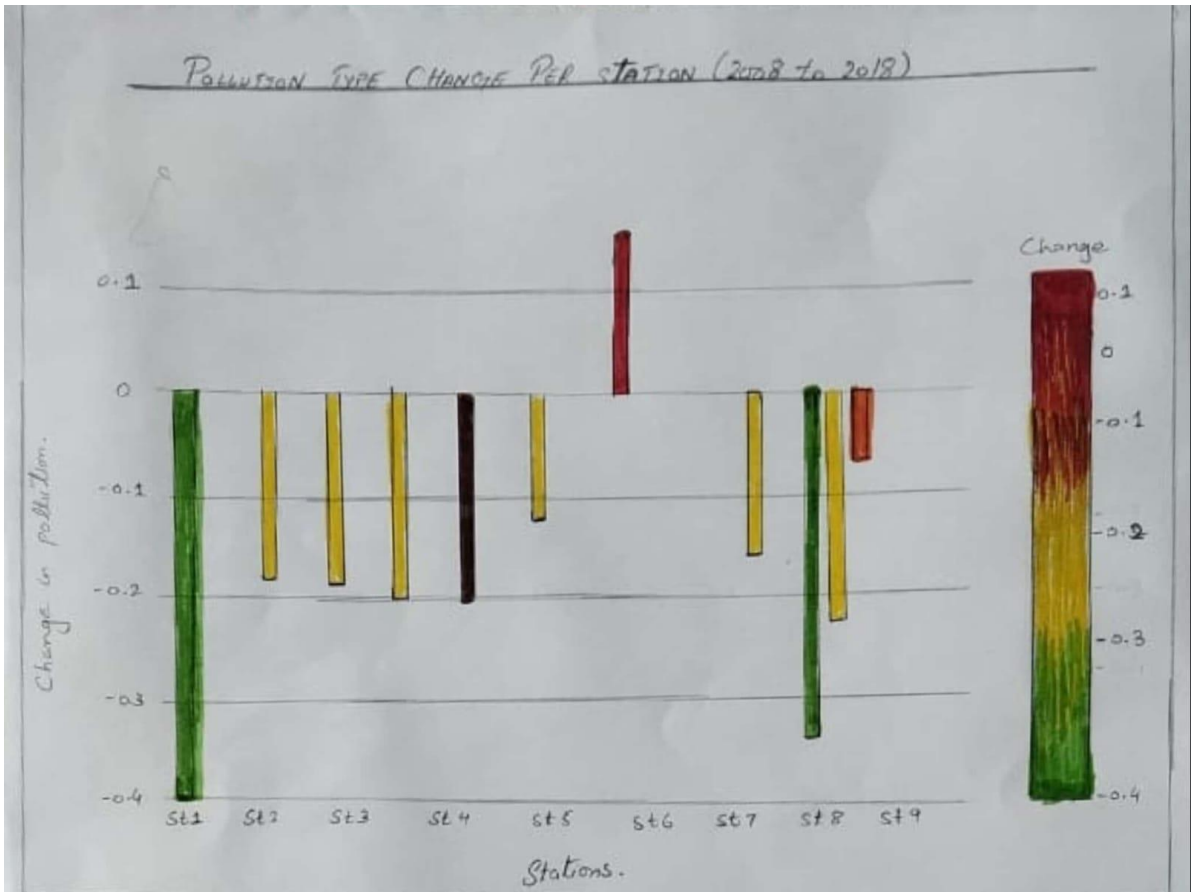


Figure 1

The above figure shows the pollution change in 2018 as compared to 2008 for each station. Green color shows decrease over time and red shows increase and the length of the bar shows the rate of change in pollution. Pollution here is a composite matrix, though the analysis will also be performed on individual matrix and per pollutants type.

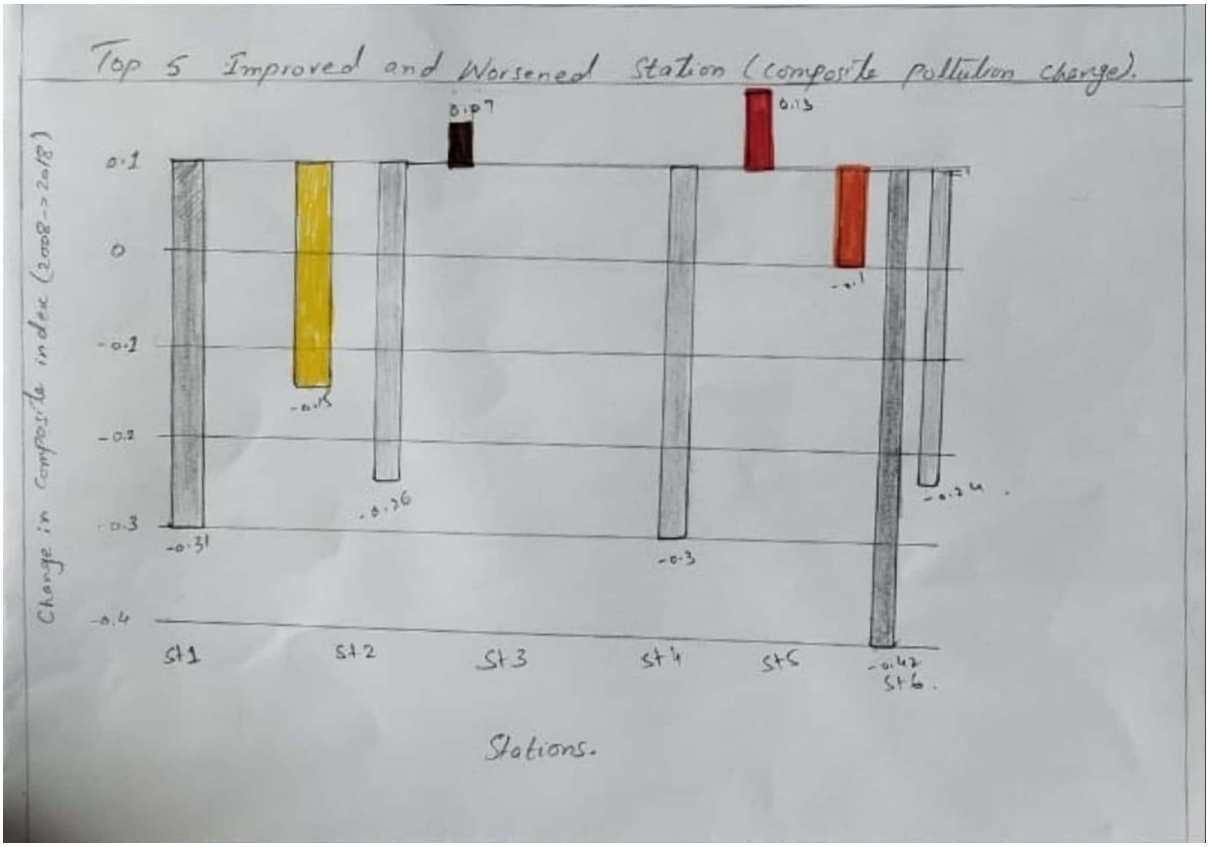


Figure 2

The above graph shows top five improved and worsened station in 2018 as compared to 2008.

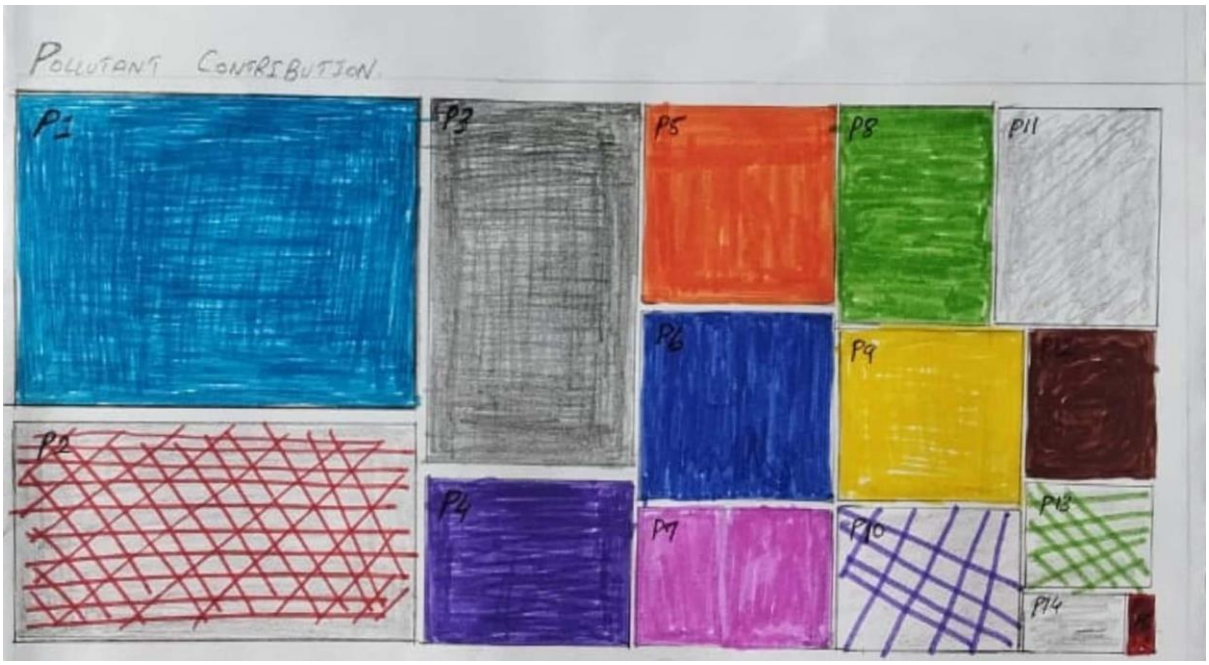


Figure 3

The above graph is a tree map which shows the overall contribution of each pollutant type in Madrid city's pollution.

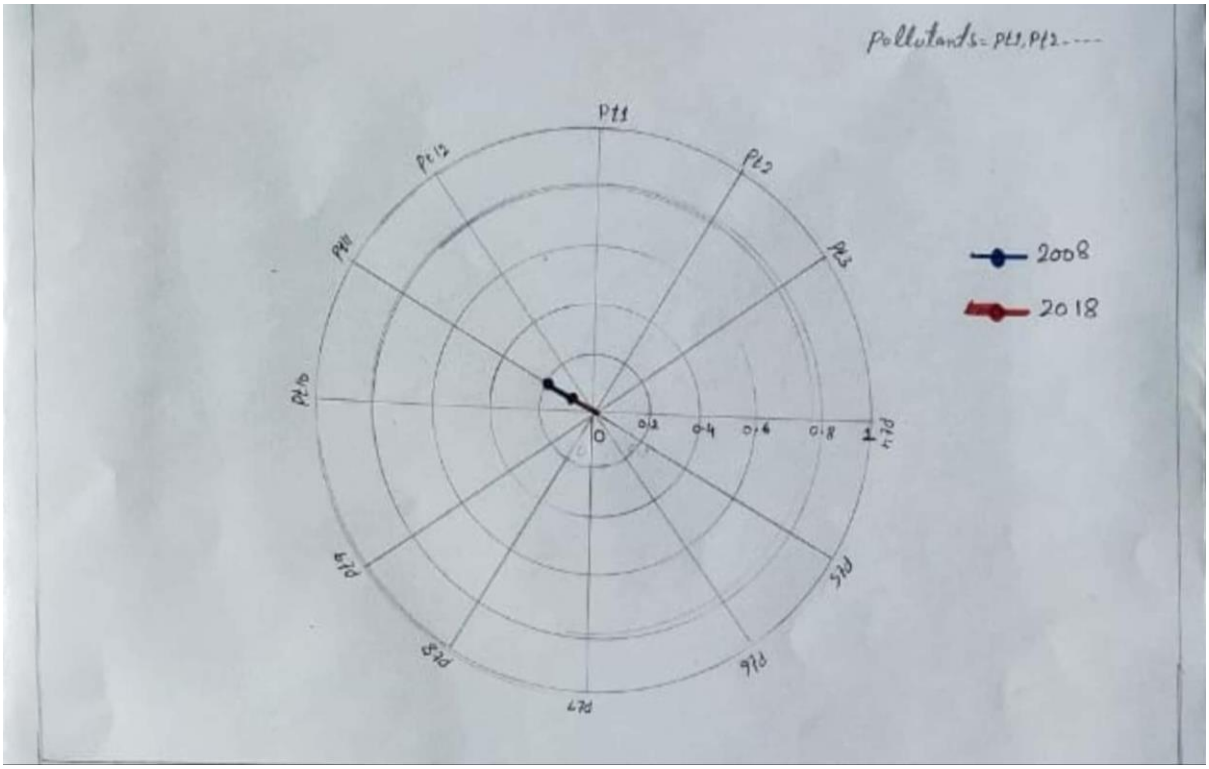


Figure 4

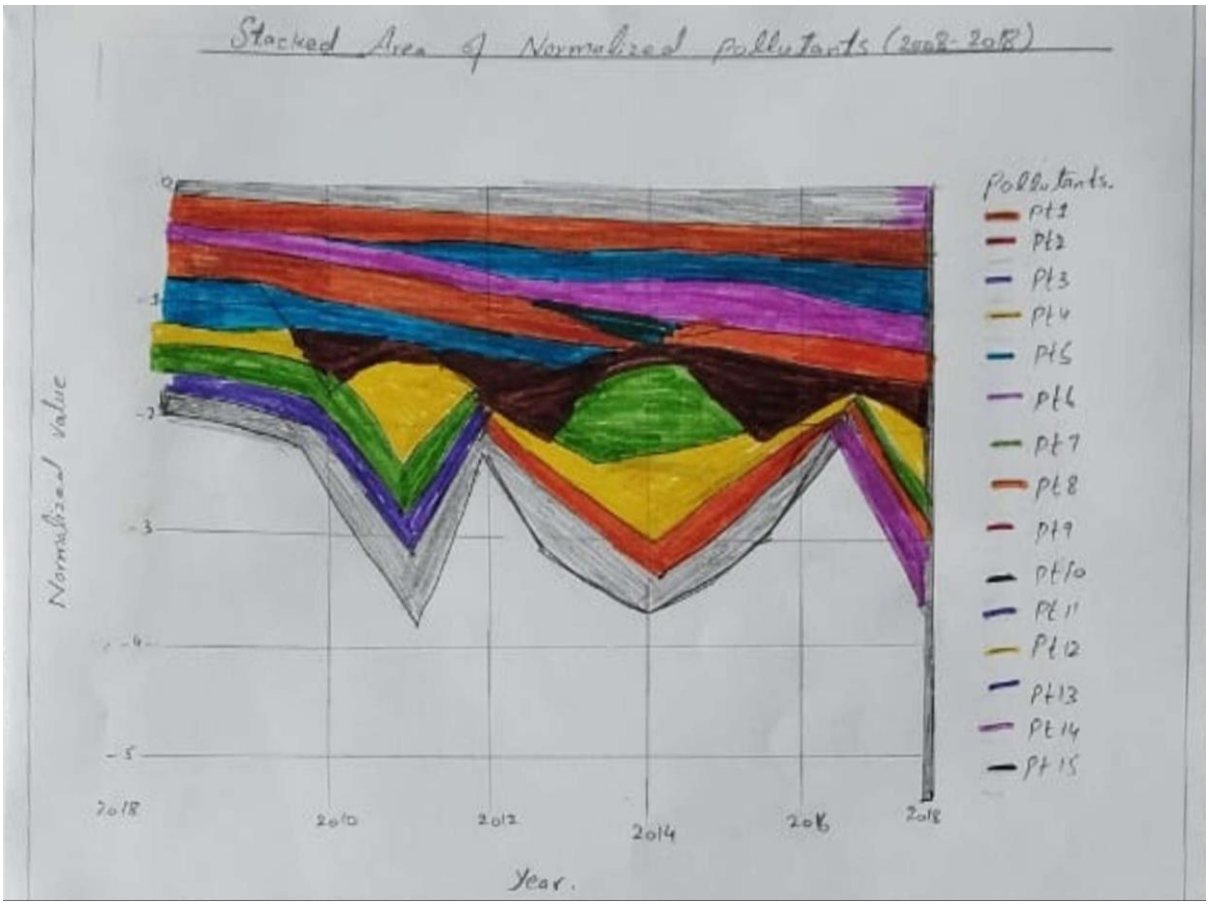


Figure 5

Figure 4 & 5 shows how different pollutant types evolved between 2008 and 2018.

Figure 6

The above line chart shows the relation between pollutant types and seasonality factor.

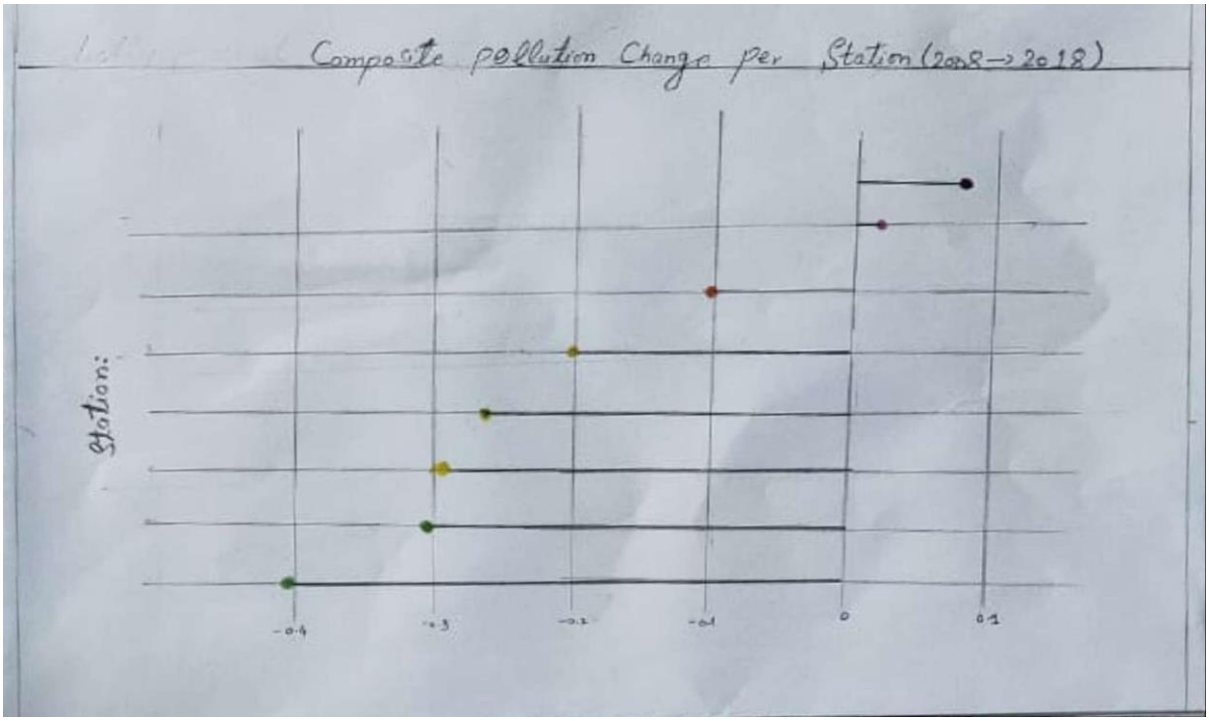
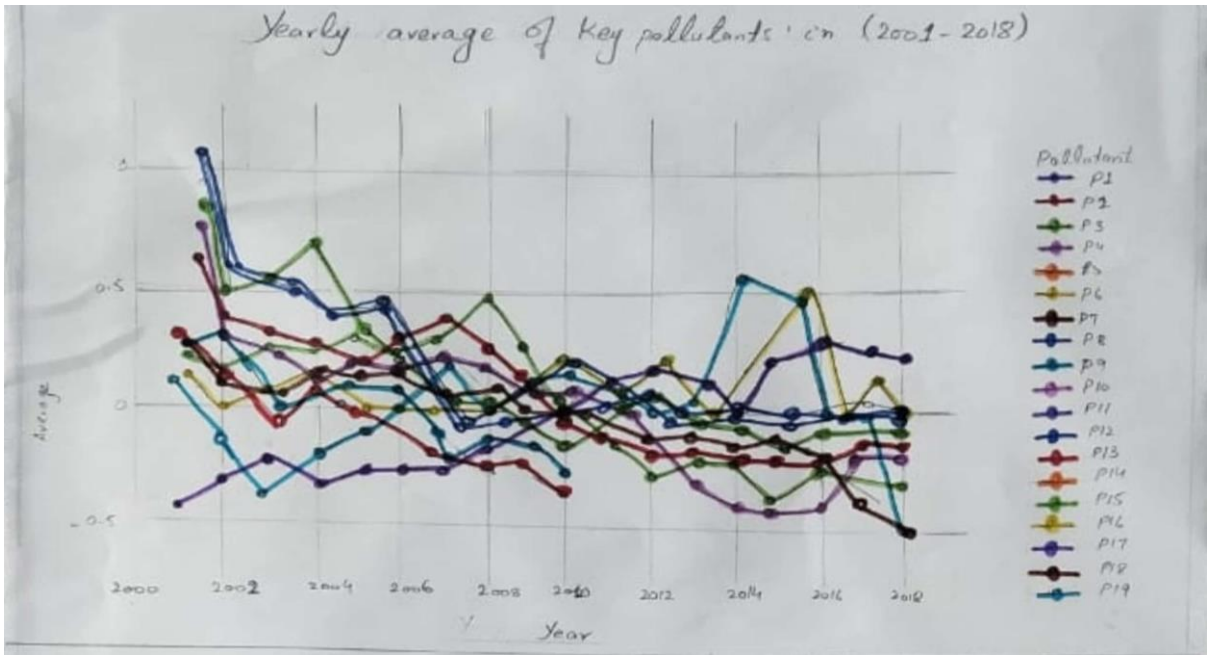


Figure 7

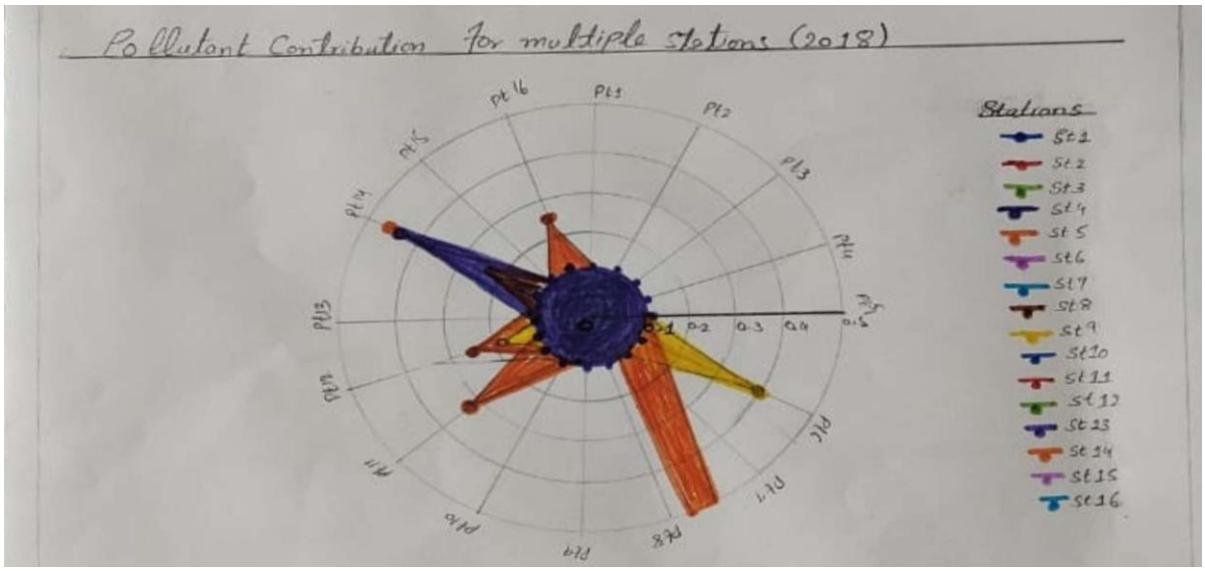
The above lolipop chart shows pollution change against each station between 2008 and 2018.

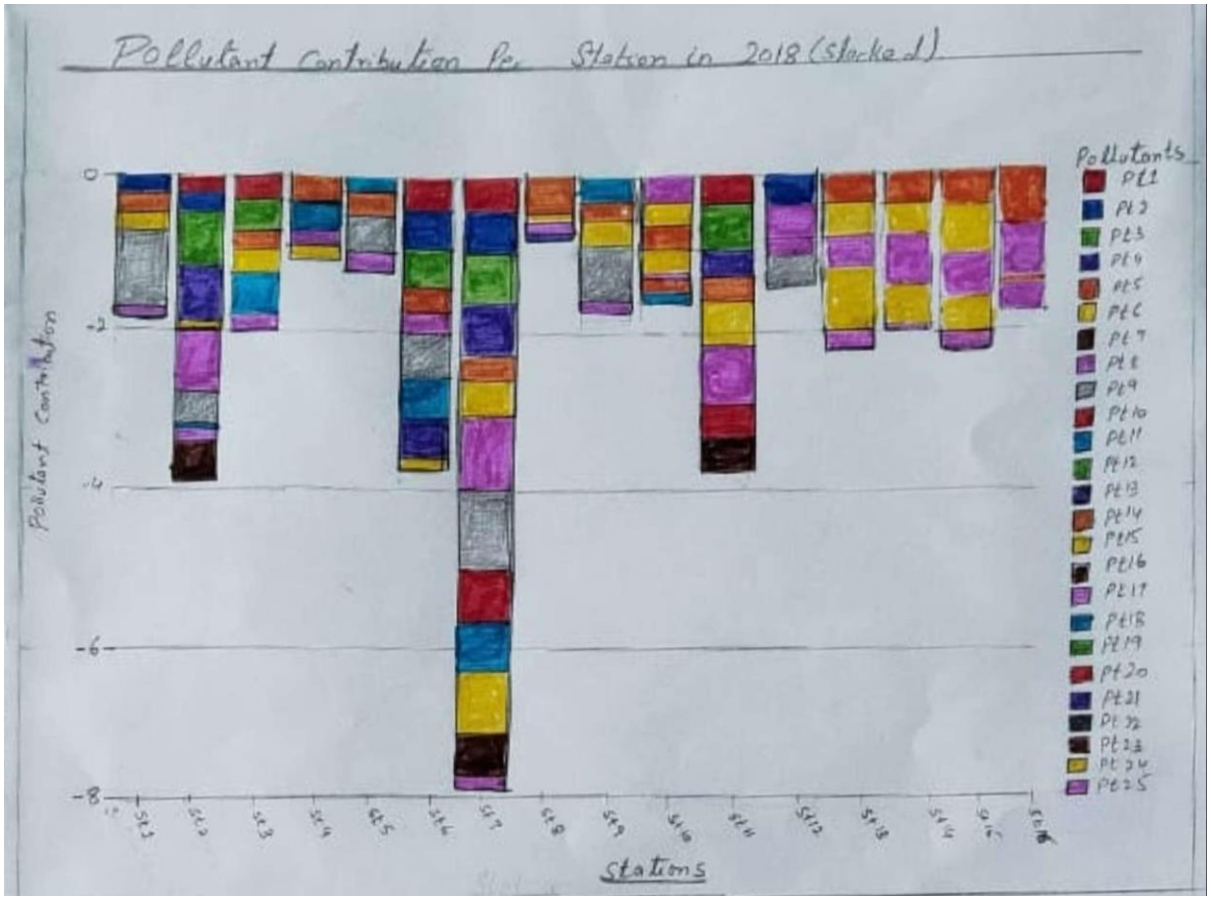
3.3 The final design

From all the sketches, we chose a few that best matched the project goals. These included line & bar charts, tree & radar maps, area plots and heatmaps — all aimed at making it easier to spot patterns and compare pollutants across time and space. They're designed to support exploration, not just give one answer.



The graph above shows changes in individual pollutant types from 2001 to 2018. It provides a compact overview of the entire dataset. The x-axis represents the years from 2001 to 2018, while the y-axis indicates the yearly average for each pollutant type. Each colored line corresponds to a different pollutant. Some lines are incomplete or missing, suggesting that data for certain years may be unavailable for specific pollutant types.





Both above graphs radar plot and stacked plot shows in each station in Madrid how much each pollutant type is in 2018.

Part 4. Implementation

In this section, we explain the visualisation implementations that were created to answer the research questions. For each visual, we describe what was originally intended, what was actually implemented using Python, and what types of interactions are available to support data exploration and interpretation.

4.1 Monthly Average of Pollutants Over Years (Animated)

4.1.1 Intended Design

The idea was to build a multi-line animated chart showing how pollution levels vary by month for each pollutant across the years. This would highlight seasonal patterns — for example, higher NO_2 and CO levels in winter — and allow comparison across multiple pollutants. The animation per year (2008–2018) was included to explore how these patterns evolve over time. This was a key visual to answer our question: *"How have the different measurements of pollution evolved between 2008 and 2018?"*

4.1.2 Actual Design

The visualisation was created using Plotly Express. Each line represents a pollutant, with months on the x-axis and scaled average concentration on the y-axis. The chart animates yearly frames, so users can observe changes from 2008 to 2018. The lines update automatically, offering a clear view of seasonal trends — such as rising O_3 levels in summer and NO_2 peaking in winter.

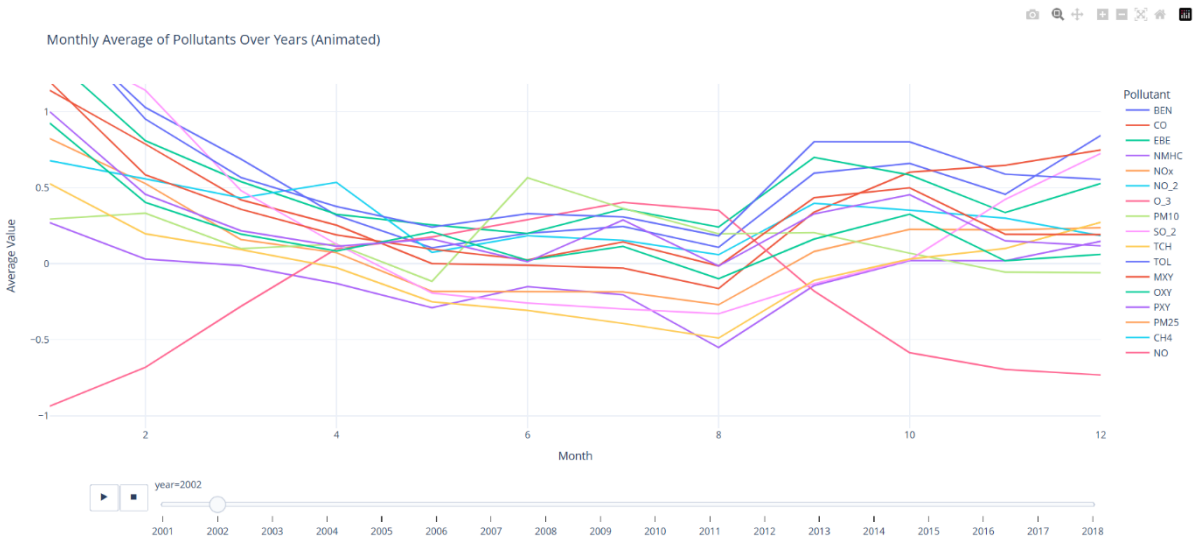


Figure 11

4.1.3 Interaction Description

- **Hover tooltips** showing month, pollutant, and value
- **Animation controls** to play/pause and scroll through years
- **Legend toggles** to show/hide specific pollutants
- **Zoom & pan** for focusing on specific time periods

4.2 Pollutants Over Time – All Stations Combined (Interactive)

4.2.1 Intended Design

This chart was originally designed to visualize how air pollution trends have changed over time for key pollutants in Madrid. The initial sketch proposed a multi-line chart, with each line representing a pollutant, plotted against years. The idea was to track overall air quality evolution from 2008 to 2018, while allowing users to compare pollutant behavior side by side. This visual directly supports our research question: *"How has pollution evolved in Madrid from 2008 to 2018?"*

4.2.2 Actual Design

The implemented chart was built using Plotly and displays a normalized time series for multiple pollutants. The x-axis represents years from 2008 to 2018, and the y-axis shows the standardized average concentration for each pollutant. Each pollutant is color-coded and appears as a distinct line. Normalization allows fair comparison, even though pollutants are measured in different units. The visual clearly highlights trends — for example, NO_2 and PM10 show a downward trend, while O_3 remains stable or slightly increases.

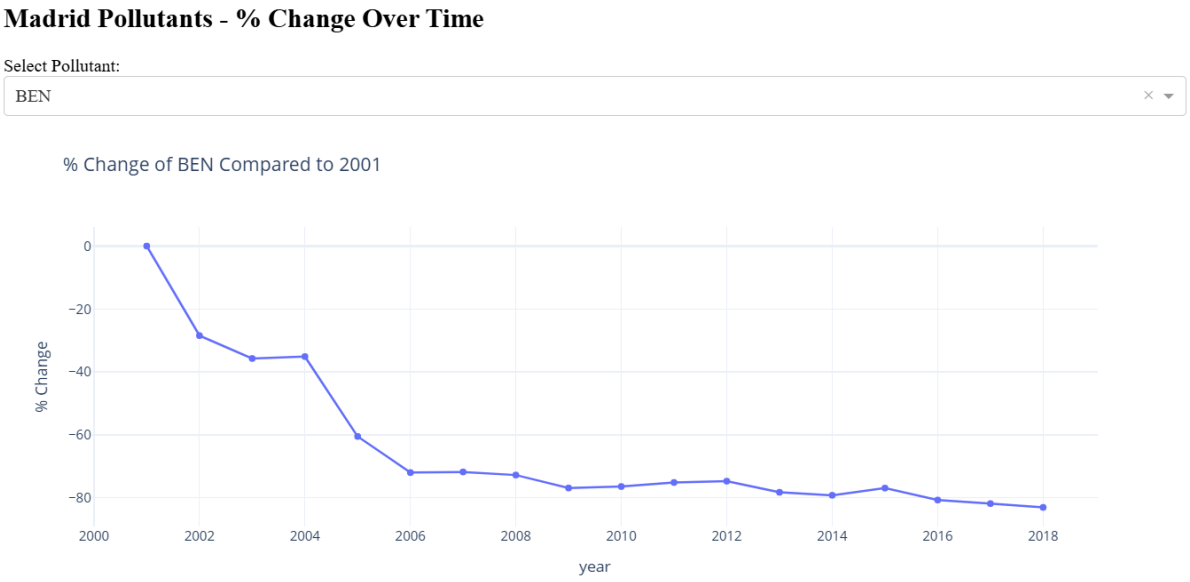


Figure 12

4.2.3 Interaction Description

- **Hover tooltips:** Display year, pollutant name, and normalized value
- **Legend click-to-toggle:** Hide/show specific pollutants to declutter the view
- **Zoom and pan:** Focus on a subset of years or steep trend changes
- **Responsive layout:** Adjusts well to screen resizing and allows seamless comparison

4.3 Pollutant Contribution per Station in 2018 (Stacked Bar Chart)

4.3.1 Intended Design

The original concept was to create a visual comparison of air pollution composition across different monitoring stations in Madrid for the year 2018. We sketched this as a stacked bar chart where each station is a bar, and each segment of the bar represents a different pollutant. The goal was to answer the question: *"Which areas were the most and least polluted in 2018, and by what pollutant?"*
This visual was meant to highlight stations with high total pollution, as well as stations with specific pollutants dominating the air composition.

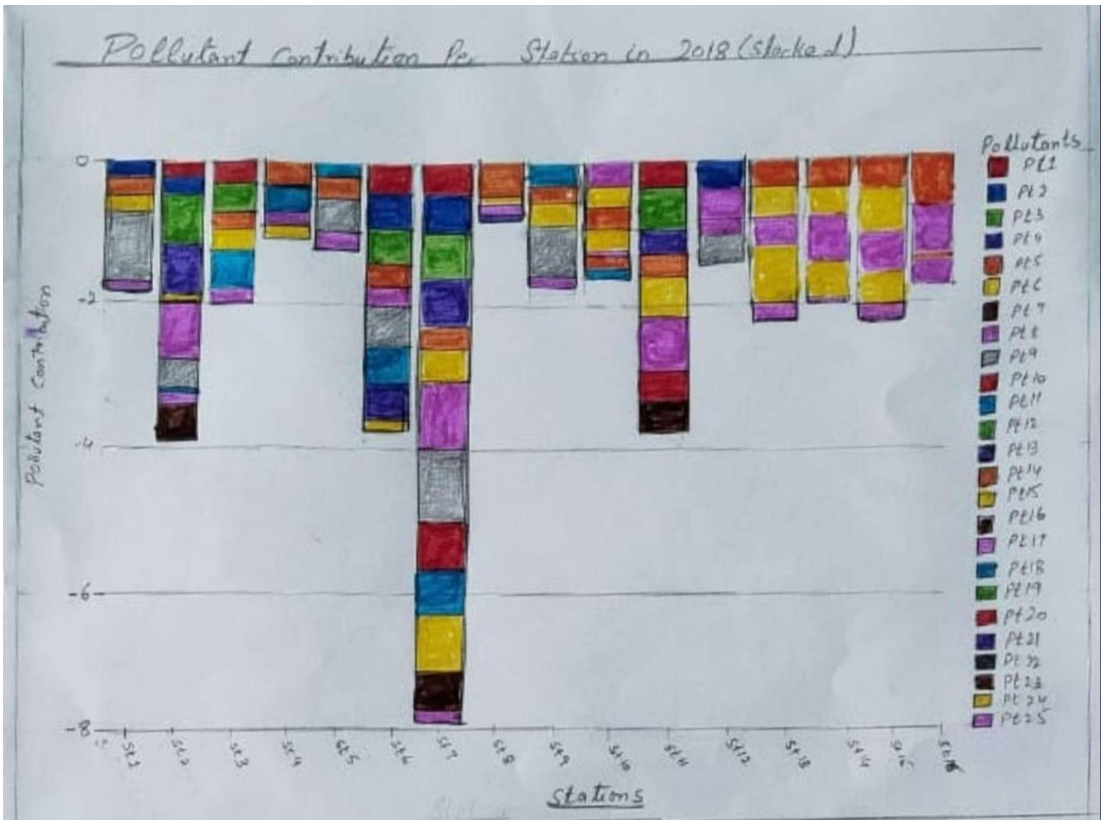


Figure 13

4.3.2 Actual Design

The final visualisation was implemented using Plotly as a horizontal stacked bar chart.

- Each bar corresponds to a monitoring station in Madrid.
- The length of the bar represents the total pollution level.
- Each segment within the bar shows the relative contribution of a pollutant (e.g., NO_2, PM10, CO, O_3, etc.).

The pollutants were normalized to allow direct comparison, and the color-coded stacks clearly show which pollutants were more prominent at which stations. The chart reveals that stations like Farolillo and Casa de Campo show significant pollution levels, while others like Escuelas Aguirre show a more balanced or moderate profile.

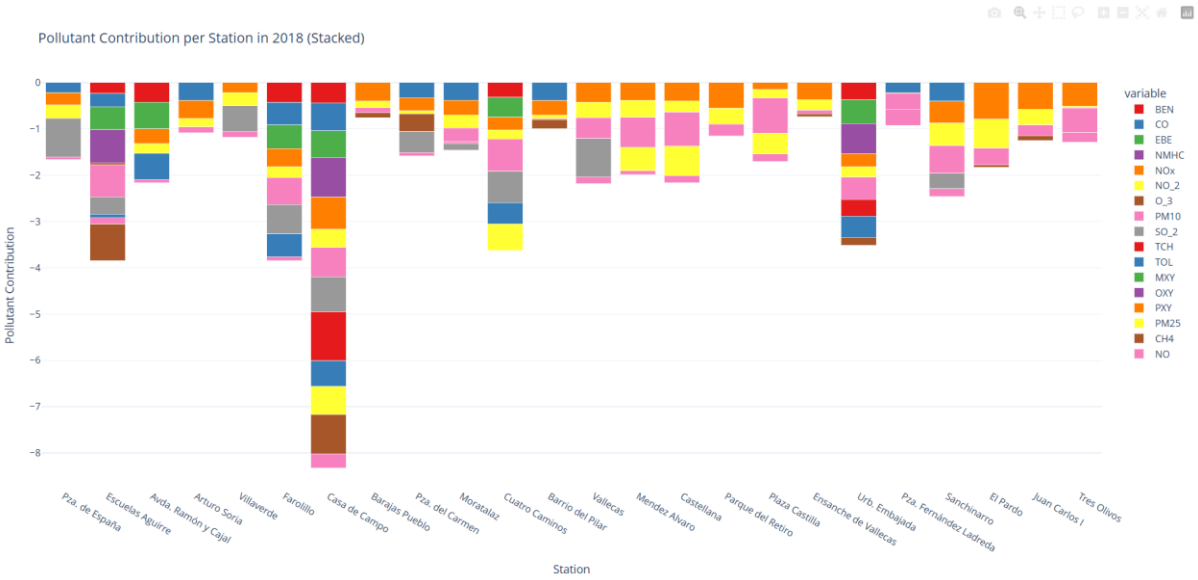


Figure 14

4.3.3 Interaction Description

- **Hover tooltips:** Display station name, pollutant type, and corresponding value for that pollutant.
- **Legend toggles:** Clicking on a pollutant name in the legend will hide or isolate that pollutant across all stations, helping reduce clutter and focus the analysis.
- **Zoom/pan:** Enables closer inspection of tightly grouped stations or very small pollutant contributions.
- **Responsive layout:** Automatically adapts to screen/window size, making the chart easy to use in presentations or reports.

4.4 Composite Pollution Change Map (2008 → 2018)

4.4.1 Intended Design

The original idea was to create a map-based visualisation that shows how overall pollution changed at each monitoring station in Madrid over the 10-year period from 2008 to 2018. In our sketches, we planned to display color-coded markers where each marker's color and size would represent the magnitude and direction of change in composite pollution (i.e., combined impact of multiple pollutants). The aim was to visually highlight which areas improved in air quality and which worsened, helping answer the question: "Which areas of Madrid improved or worsened in pollution levels between 2008 and 2018?"

4.4.2 Actual Design

Implemented using Plotly Mapbox, each marker on the map represents a station.

- **Color:** Green for improvement, red for worsening
- **Size:** Indicates how much pollution changed
- **Location:** Based on real latitude and longitude of each station

This combines multiple pollutants into a single "composite change" score, offering a clear spatial summary of trends from 2008 to 2018.

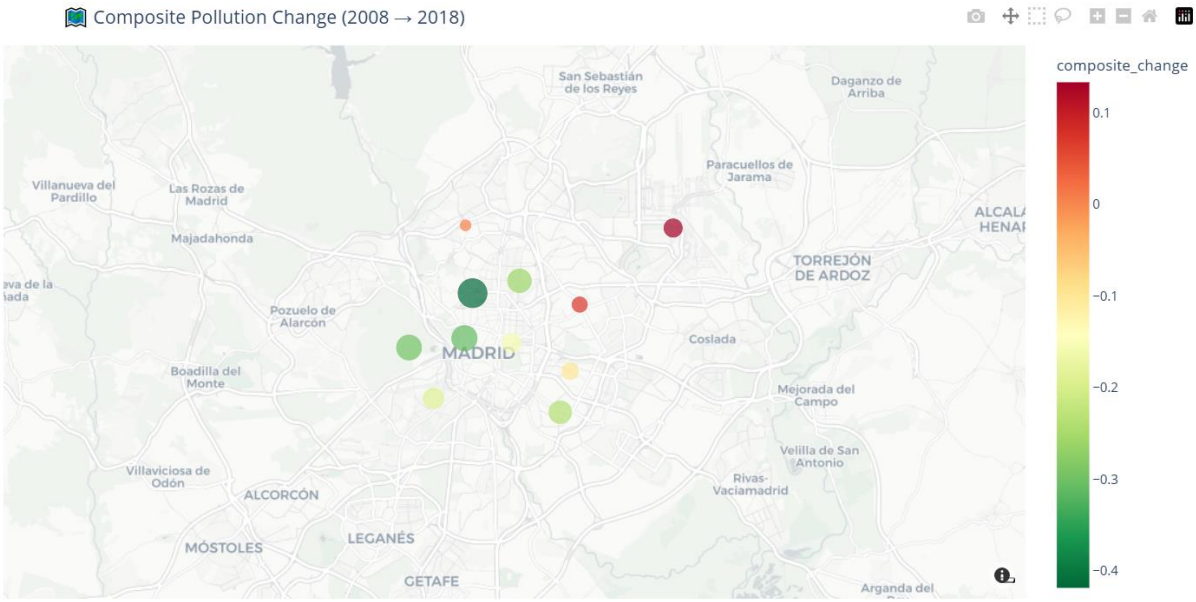


Figure 15

4.4.3 Interaction Description

- **Hover tooltips** show station name and change value
- **Zoom/pan** lets users explore different regions
- **Color legend** explains direction and severity of change

These interactions help users explore where pollution improved or worsened most in the city.

4.5 Composite Pollution Change per Station (Lollipop Chart)

4.5.1 Intended Design

The goal of this visualisation was to clearly show how much the pollution level at each station changed between 2008 and 2018. We originally planned to use a lollipop chart, where each station would be a horizontal line ending in a circle, and the length and direction of the line would represent the magnitude and direction of change. This was intended to make it easy to spot which stations had improved (negative values) or worsened (positive values) in overall pollution, based on a composite score combining several pollutants.

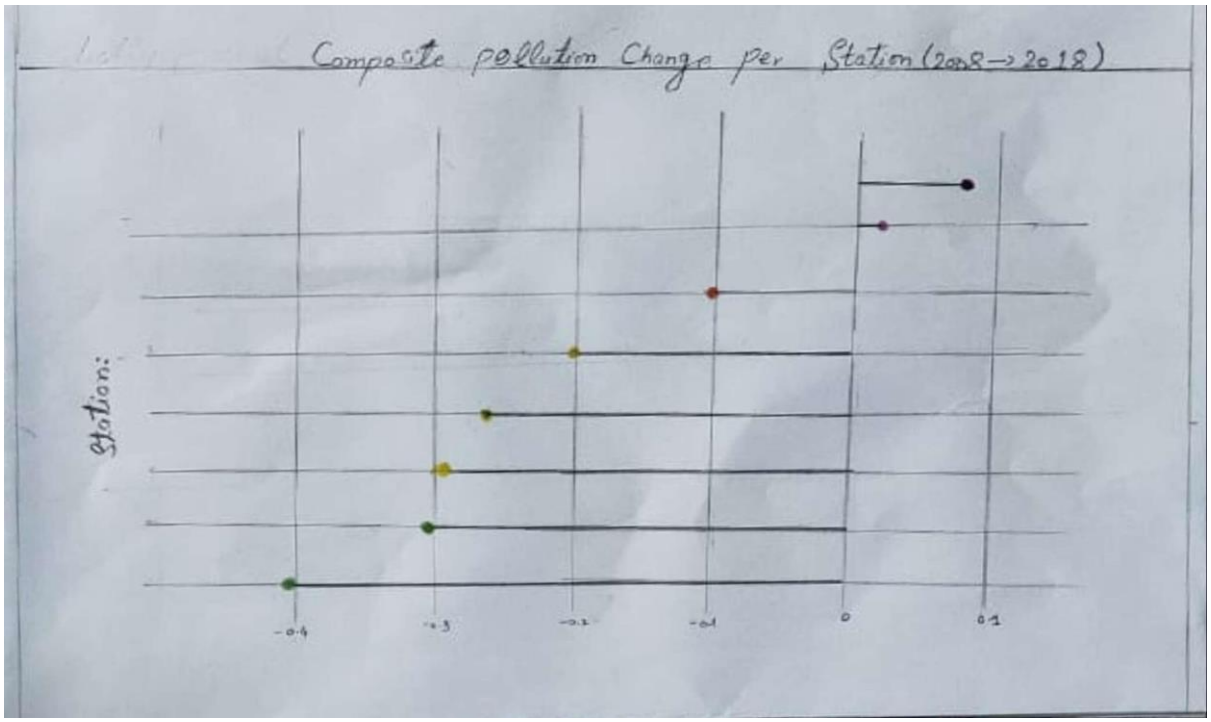


Figure 16

4.5.2 Actual Design

The chart was implemented using Plotly as an interactive horizontal lollipop plot.

- **Y-axis:** Station names
- **X-axis:** Composite pollution change score (2018 minus 2008)
- **Line and Dot:** Each station has a line from 0 to its change value, ending in a colored circle
- **Color:** Red indicates an increase in pollution, green indicates a decrease

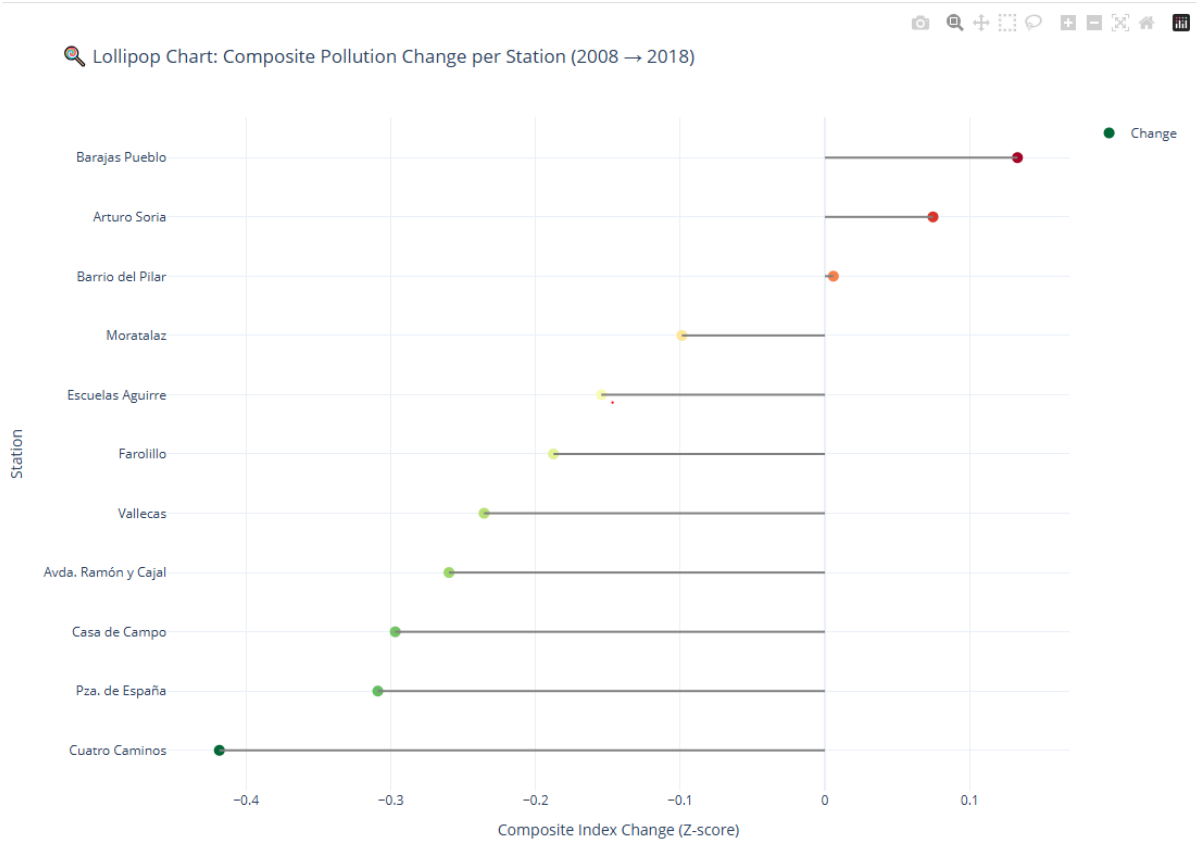


Figure 17

4.5.3 Interaction Description

- **Hover tooltips** show station name and exact change value
- **Scroll support** allows smooth viewing of all stations, even if many are plotted
- **Color coding** helps instantly distinguish between positive and negative change
- **Responsive layout** ensures readability across devices and screen sizes

4.6 Normalized Trend of Pollutants (2008–2018)

4.6.1 Intended Design

The goal was to create a visualisation that compares long-term pollution trends across different pollutants in Madrid from 2008 to 2018. Since pollutants are measured in different units, we proposed using normalized values so that all lines could be plotted on the same scale. The original sketch envisioned a multi-line plot, where each line represents a pollutant, making it easy to observe which pollutants decreased, increased, or stayed constant over time.

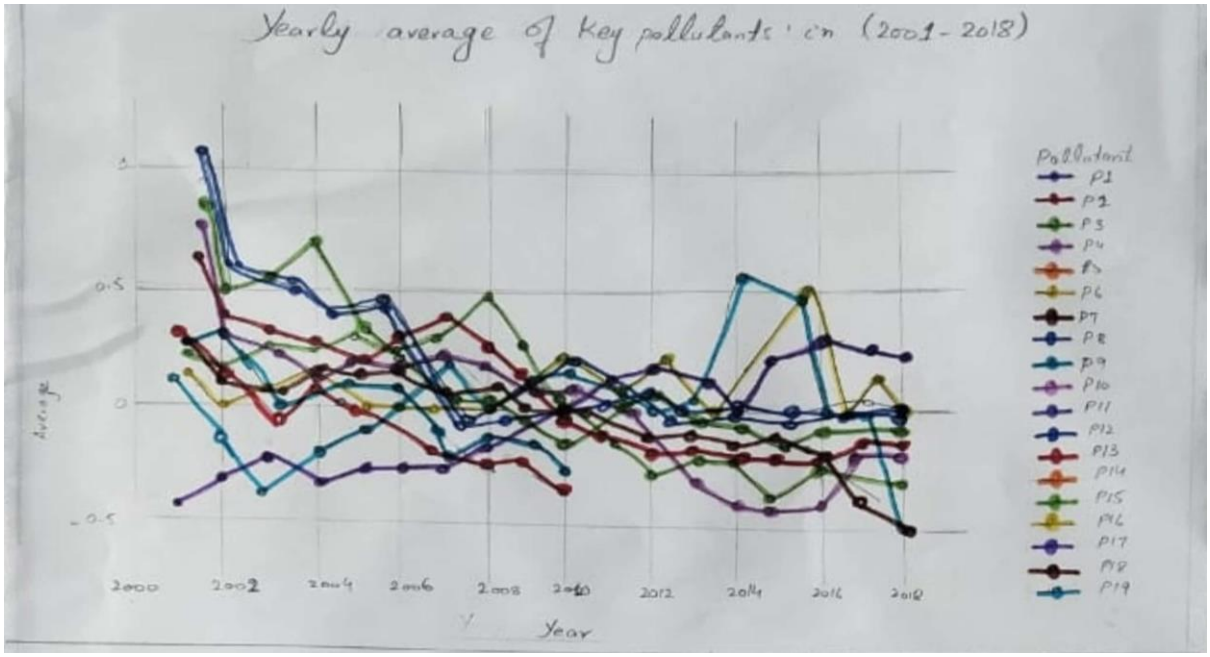


Figure 18

4.6.2 Actual Design

This chart was implemented using Plotly as a fully interactive line chart.

- **X-axis:** Years (2008 to 2018)
- **Y-axis:** Normalized pollutant levels (scaled)
- **Lines:** Each pollutant is represented as a colored line
- **Normalization:** Allows fair comparison between pollutants with different units

The chart highlights that while pollutants like NO_2 and PM10 have generally decreased, others like O_3 have remained more stable or slightly increased.

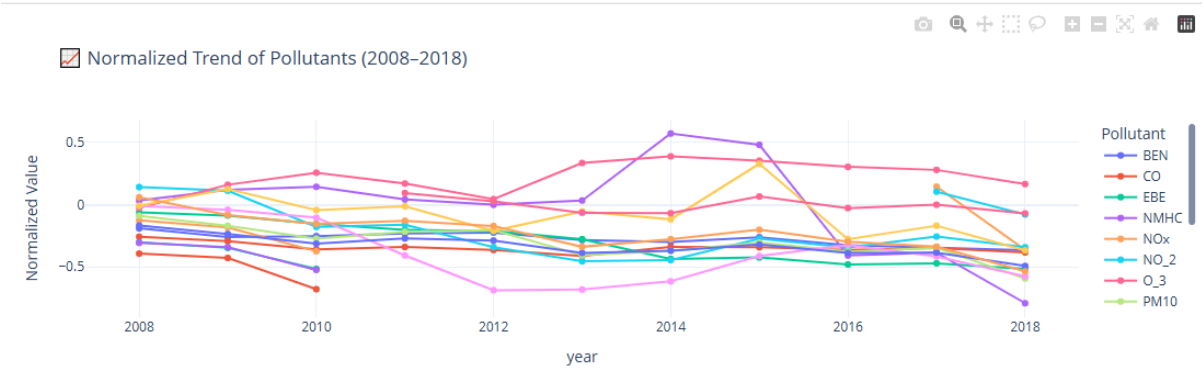


Figure 19

4.6.3 Interaction Description

- **Hover tooltips** display the pollutant name, year, and normalized value
- **Zoom and pan** enable users to focus on specific time periods
- **Legend toggling** allows users to hide or isolate specific pollutants
- **Box/lasso** selection helps highlight subsets of the data for detailed analysis

Part 5. Links

GitHub Repository: https://github.com/sadiazaman-git/Madrid-Air-Pollution-Analysis

YouTube Demonstration: [Insert YouTube Video URL here]