Data Curation Report

In this report, I outline the steps taken to curate the raw data provided in the CSV files for the five visualisations. The primary goal of the data curation process is to prepare a clean and consistent dataset for further analysis and visualisation.

1. Nitric Oxide, PM10, and Nitrogen oxides as nitrogen dioxide Data Curation:

   Columns Affected: Date, time, Nitric Oxide, PM₁₀ particulate matter (Hourly measured), Nitrogen oxides as nitrogen dioxide

   Data Import: The CSV files for the years 2018 to 2023 were imported using the read_csv function from the ‘readr’ package. Data import started from the 5th row to skip any header information.

   Filtering by Date: The data was filtered based on specific dates of interest for each year using the filter function from the ‘dplyr’ package. Only rows corresponding to the specified dates were retained.

   Handling Missing Values: Rows containing missing values were identified and removed using the ‘na.omit’ function to ensure only complete observations were included.

   Data Formatting: The date column was converted from a character format to a Date format using the as.Date function to ensure consistency.

   Combining Data: Finally, the filtered and cleaned data from each year were combined into a single tibble using the bind_rows function.

2. Monthly Average Data Curation:

   Columns Affected: Date, time, Nitric Oxide, PM₁₀ particulate matter (Hourly measured), Nitrogen oxides as nitrogen dioxide

   Data Import: The CSV file for the year 2020 was imported, and relevant columns were selected using the read_csv and select functions.

   Data Transformation: Monthly average values for PM10 particulate matter, Nitric oxide, and Nitrogen oxides as nitrogen dioxide were calculated using the mutate, group_by, and summarise functions.

   Data Reshaping: The data was reshaped into long format using the pivot_longer function to facilitate visualisation.

3. Fifth Analysis Visualisation:
   Description: Comparison of the air pollutant concentration before and after Clean Air Zone was Implemented. The PM₁₀ particulate matter (Hourly measured) Column was excluded because of gaps of data in the excel file. This would prevent the results from being skewed.

   Columns Affected: Date, time, Nitric Oxide, Nitrogen dioxide, Nitrogen oxides as nitrogen dioxide

   Data Import: CSV files for the years 2020 to 2022 were imported, and relevant columns were selected.

   Data Processing: The Date column was converted to the Date type, and data from all years were combined into a single dataset.

   Data Filtering: Data was filtered for two periods: before and after the implementation of the Clean Air Zone (CAZ) on 29th November 2021. It excluded 29th November 2021.

   Data Transformation: Daily averages for Nitrogen dioxide, Nitric oxide, and Nitrogen oxides as nitrogen dioxide were calculated for both periods using the group_by and summarise functions.

An example of a data inconsistency problem is shown below;

29/11/2021	01:00				1.3693	R	ugm-3	9.09626	R	ugm-3
29/11/2021	02:00				1.28095	R	ugm-3	9.23172	R	ugm-3
29/11/2021	03:00				0.76195	R	ugm-3	6.46521	R	ugm-3
29/11/2021	04:00				0.96072	R	ugm-3	8.29782	R	ugm-3
29/11/2021	05:00				1.15948	R	ugm-3	10.33399	R	ugm-3
29/11/2021	06:00				2.71651	R	ugm-3	11.76335	R	ugm-3
29/11/2021	07:00				15.20581	R	ugm-3	24.26699	R	ugm-3
29/11/2021	08:00				21.26825	R	ugm-3	32.63022	R	ugm-3
29/11/2021	09:00				36.44093	R	ugm-3	47.27889	R	ugm-3
29/11/2021	10:00				24.18353	R	ugm-3	37.82932	R	ugm-3
29/11/2021	11:00				29.02023	R	ugm-3	37.07974	R	ugm-3
29/11/2021	12:00				36.70596	R	ugm-3	40.13807	R	ugm-3
29/11/2021	13:00	39.614	R	ugm-3 (Ref	29.85948	R	ugm-3	39.23648	R	ugm-3
29/11/2021	14:00	24.155	R	ugm-3 (Ref	49.89095	R	ugm-3	53.78023	R	ugm-3
29/11/2021	15:00	26.087	R	ugm-3 (Ref	42.00646	R	ugm-3	58.49056	R	ugm-3
29/11/2021	16:00	21.256	R	ugm-3 (Ref	63.80476	R	ugm-3	60.07934	R	ugm-3

After cleaning the data;

21	03/03/2021	21:00:00	66.668
22	03/03/2021	22:00:00	67.634
23	03/03/2021	23:00:00	72.465
24	03/03/2021	24:00:00	68.600
25	29/11/2021	13:00:00	39.614
26	29/11/2021	14:00:00	24.155
27	29/11/2021	15:00:00	26.087
28	29/11/2021	16:00:00	21.256
29	29/11/2021	17:00:00	22.223
30	29/11/2021	18:00:00	19.324
31	29/11/2021	19:00:00	26.087
32	29/11/2021	20:00:00	11.594
33	29/11/2021	21:00:00	7.730
34	29/11/2021	22:00:00	9.662
35	29/11/2021	23:00:00	8.696

All the missing data for 29/11/2021 are omitted.