IJC 437 Introduction to Data Science 2025-2026

Air Quality Dynamics in Sheffield and Hull: A Data Science Approach to Environmental Analysis

Student Registration Number: 250124882

Word count: 2002

January 9, 2026

Abstract

Air quality depends on both local emissions and larger atmospheric processes, which create different pollution patterns in each city. This study looks at air quality in Sheffield and Hull from 2023 to 2025, using exploratory data analysis to compare how NO_2, O_3, $PM_{2.5}$, and PM_{10} change over time in these two cities. The research uses open-source hourly monitoring data, which were cleaned and grouped into daily, monthly, and seasonal sets. Methods such as Locally Estimated Scatterplot Smoothing (LOESS), Seasonal Trend Decomposition using LOESS (STL), and correlation analysis were used to examine short-term changes, seasonal cycles, and long-term trends.

The results show that Sheffield has higher levels of NO_2 and $PM_{2.5}$, which matches its heavy traffic emissions. Hull, on the other hand, has higher PM_{10} and O_3 levels, likely due to its industrial and coastal location. Both cities have clear seasonal patterns; traffic-related pollutants peak in winter, while ozone peaks in spring and summer. Over time, NO_2 levels have declined, and O_3 levels have increased, likely due to changes in emissions. This study focuses on understanding and comparing these patterns over different time periods, rather than making predictions.

Chapter 1

Introduction

1.1 Background and Context

Air pollution remains a major environmental and public health concern in the United Kingdom, contributing to significant illness and premature mortality each year (OHID, 2022). Pollutants such as nitrogen dioxide (NO_2), ozone (O_3), and particulate matter ($PM_{2.5}$ and PM_{10}) are strongly associated with cardiovascular and respiratory disease, making them a central focus of national clean air strategies (DEFRA, 2023). Despite long-term efforts to reduce emissions, air quality still varies across cities. These differences are influenced by local emission sources, urban layout, geography, and weather conditions.

Sheffield and Hull provide a useful comparative case. Sheffield is an inland, densely populated city where air quality is strongly influenced by road traffic and residual industrial activity (Sheffield City Council, 2010). Hull, in contrast, is a coastal port city affected by maritime transport, petrochemical industries, and enhanced atmospheric dispersion (Hull City Council, 2023). These contrasting environments make the two cities suitable case studies for examining how pollution behaves across different settings and seasons. The growing availability of monitoring data now allows these patterns to be explored in greater detail than was previously possible.

1.2 Literature Review

Previous research shows that pollutant behaviour is shaped by interactions between emission sources and atmospheric processes (Gibson et al., 2024). Seasonal effects are especially pronounced, with winter conditions often leading to elevated NO_2 and particulate matter concentrations due to reduced dispersion and increased combustion, while spring and summer sunlight promote photochemical ozone formation (Carslaw & Beevers, 2005; Monks et al., 2015).

Other studies find that inland and coastal cities have different pollution patterns. Sicard et al. (2020) found that coastal areas usually have lower NO_2 but higher ozone levels due to better airflow and different chemical reactions. DEFRA (2023) also reports that cities with lots of traffic have higher NO_2 and $PM_{2.5}$, while industrial and coastal areas often see more PM_{10} and O_3.

This study adds to previous research by using data science methods to see how these patterns appear in Sheffield and Hull. The goal is to understand how pollution changes over time, not to make predictions.

1.3 Aims and Research Questions

1.3.1 Aims

To analyse and compare the temporal behaviour of NO_2, O_3, $PM_{2.5}$, and PM_{10} in Sheffield and Hull from 2023 to 2025 using exploratory data analysis techniques.

1.3.2 Research Questions

- How do key air pollutants vary between Sheffield and Hull from 2023 to 2025?
- What daily, monthly, and seasonal patterns can be observed in pollutant levels across the two cities?

Chapter 2

Methodology

2.1 Data Description and Methods

Hourly air quality data for NO_2, $PM_{2.5}$, PM_{10}, and O_3 were obtained from the OpenAQ API for Sheffield and Hull between 2023 and 2025. Data were cleaned and processed in RStudio and Microsoft Excel. Timestamps were standardised, duplicates removed, and missing values addressed before analysis. Hourly observations were aggregated into daily, monthly, and seasonal datasets, with seasons defined as Winter (Dec–Feb), Spring (Mar–May), Summer (Jun–Aug), and Autumn (Sep–Nov)

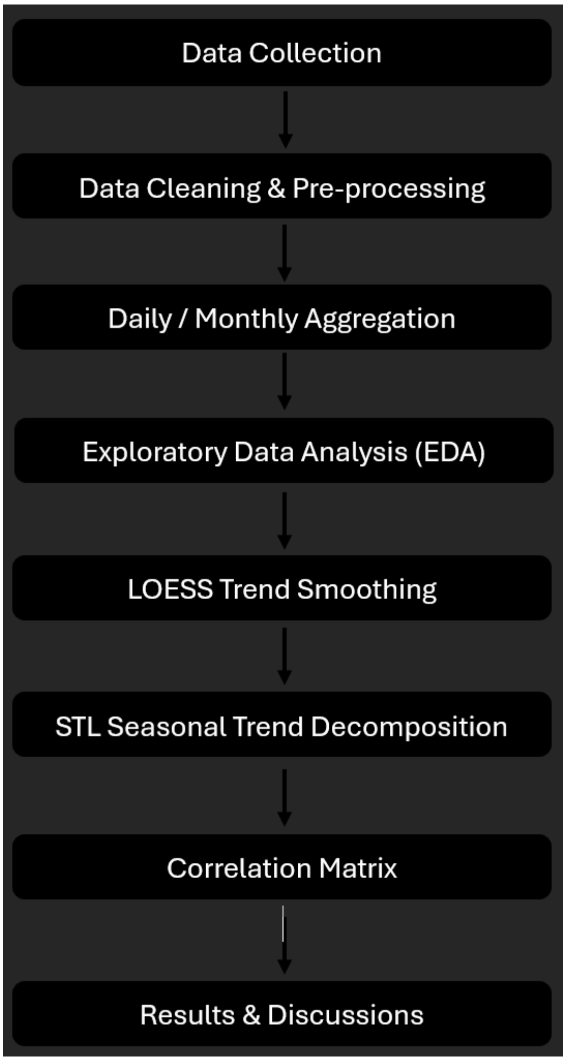
Exploratory Data Analysis (EDA) techniques were used to examine variability, seasonal differences, and relationships among pollutants. Because of high data noise, Locally Estimated Scatterplot Smoothing (LOESS) was applied to reveal medium-term trends. Seasonal Trend Decomposition using LOESS (STL) was applied to the monthly series to separate trend, seasonal, and remainder components. Correlation matrices were generated to assess relationships among pollutants and infer shared emission sources or chemical interactions.

RQ 1 is addressed by combining seasonal averages with LOESS and the long-term components extracted through STL decomposition. Together, these approaches enable comparison of overall pollutant levels between Sheffield and Hull and the identification of longer-term differences in their behaviour

RQ 2 is addressed using daily scatter plots, monthly summaries, seasonal groupings, the seasonal components from STL decomposition, and correlation analysis. This combination highlights short-term variability, recurring seasonal cycles, and the relationships between pollutants across the two cities.

All analyses used reproducible R scripts (see Appendix and GitHub profile). The study adopted a descriptive and exploratory approach because the research questions focused on interpretation rather than prediction.

Figure 2.1: Graphical Illustration of the Methodology



2.2 Analysis

The analysis used complementary steps to explore short-term variability and longer-term trends in pollutant behaviour across Sheffield and Hull. Initial analysis revealed substantial variation in the data, occasional outliers, and differences in average pollutant levels between the two cities.

Table 1.0: Daily Descriptive Statistics

City	Pollutant	N	Mean	Median	Standard Deviation	Interquartile Range
Hull	NO_2	328	16.97288	15.32124	9.943372	10.30348
	$PM_{2.5}$	328	7.559743	6.434783	3.839238	4.447908
	PM_{10}	328	14.48975	13.36364	5.720636	6.6853
	O_3	328	48.62344	48.7533	15.58634	17.73223
Sheffield	NO_2	328	15.16099	12.79724	10.0452	12.90284
	PM_{25}	328	7.045114	6.043478	3.71254	5.152174
	PM_{10}	328	12.02268	10.86957	5.300609	6.826087
	O_3	328	46.45978	46.10501	16.2674	17.77408

To make temporal patterns easier to interpret, Locally Estimated Scatterplot Smoothing (LOESS) was applied to the daily time series, revealing medium-term changes that were difficult to discern in the raw data. Seasonal Trend Decomposition using LOESS (STL) was then used to decompose each pollutant into seasonal, trend, and residual components, enabling recurring seasonal cycles to be distinguished from longer-term trends.

Relationships between pollutants were examined using correlation matrices, which showed how particulate matter, nitrogen dioxide, and ozone interact within each city's atmosphere. All analyses used reproducible R scripts (see Appendix and GitHub Profile). No predictive modelling was done, as the study focused on describing and interpreting observed patterns rather than predicting future concentrations.

Chapter 3

Results

3.1 Exploratory Data Analysis

Figure 3.1: Seasonal Average Bar Plot

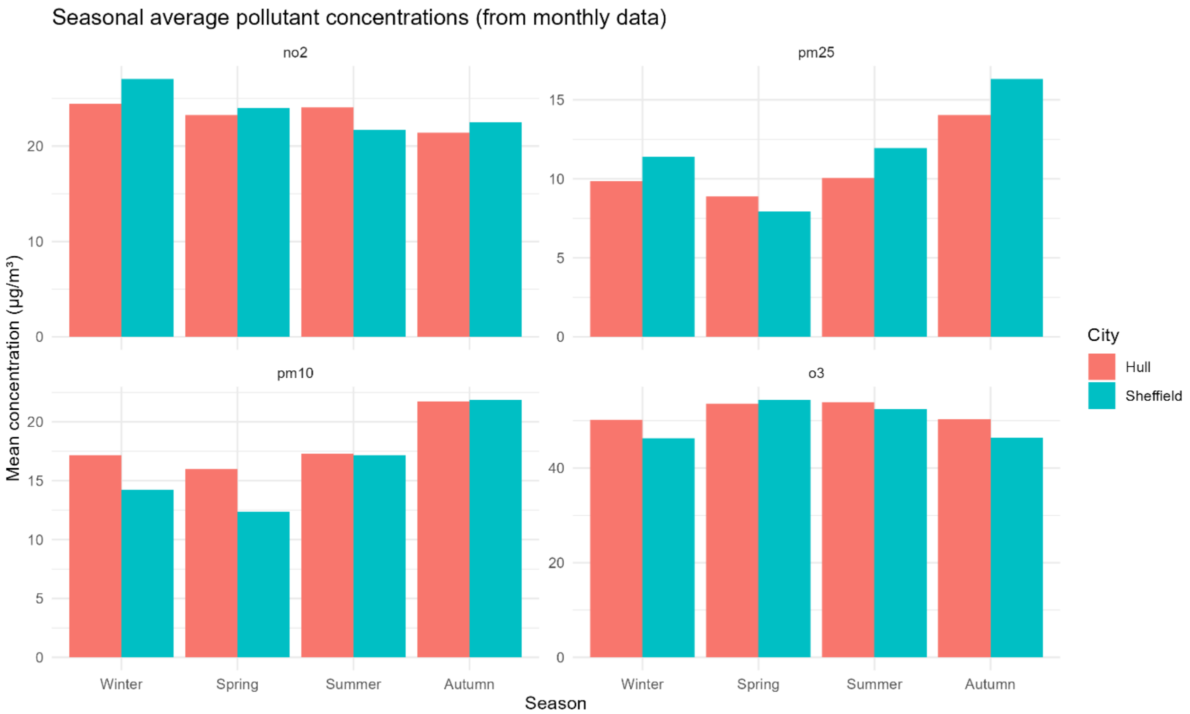


Figure 3.1 highlights clear patterns in both cities. NO_2 levels are highest in winter and lowest in summer, with Sheffield consistently showing higher values than Hull, likely due to greater traffic influence. $PM_{2\cdot5}$ and PM_{10} also rise during colder months. Sheffield has higher $PM_{2\cdot5}$, while Hull has higher PM_{10}, reflecting differences in combustion and industrial sources. Ozone peaks in spring and summer, with Hull generally recording higher values, possibly due to reduced ozone titration.

Figure 3.2: Daily Average Scatter Plot

Figure 3.2 shows substantial short-term variability and occasional pollution spikes, particularly in 2025. Although both cities exhibit similar ranges of fluctuation, the density of observations makes it difficult to identify underlying trends from the raw data alone.

3.1.1 Locally Estimated Scatterplot Smoothing (LOESS)

Figure 3.3: LOESS Trend

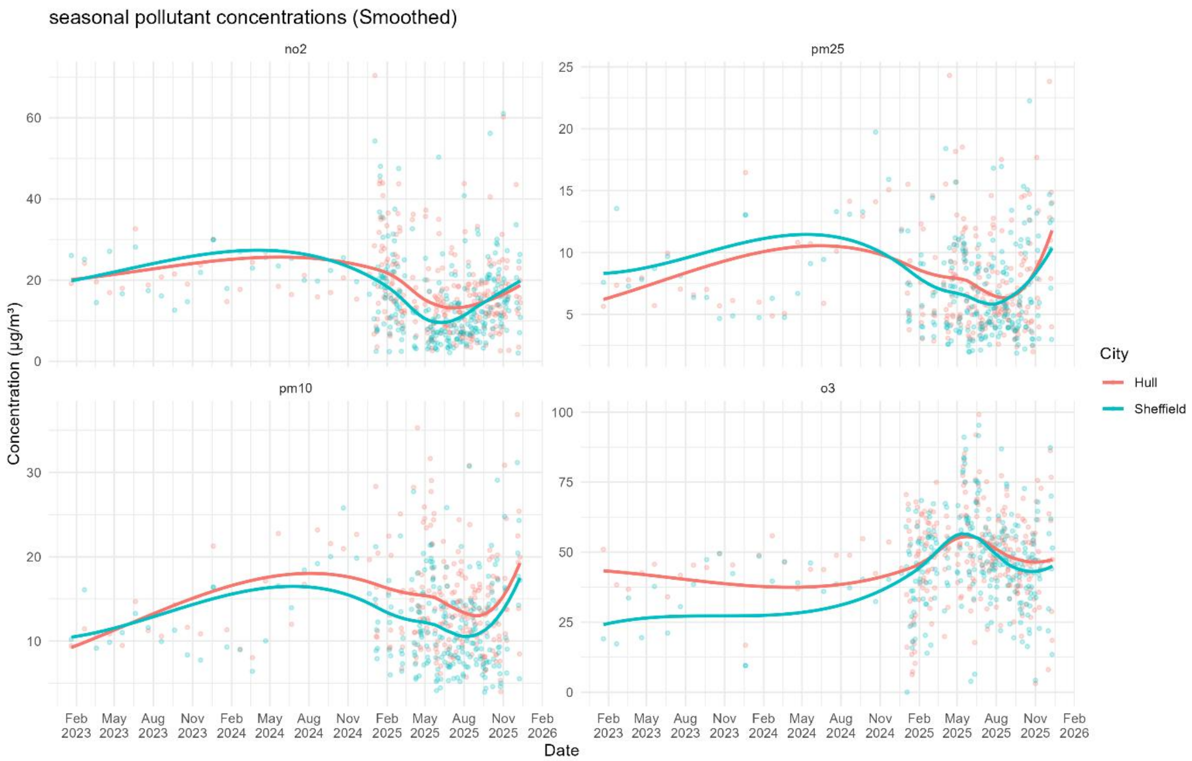
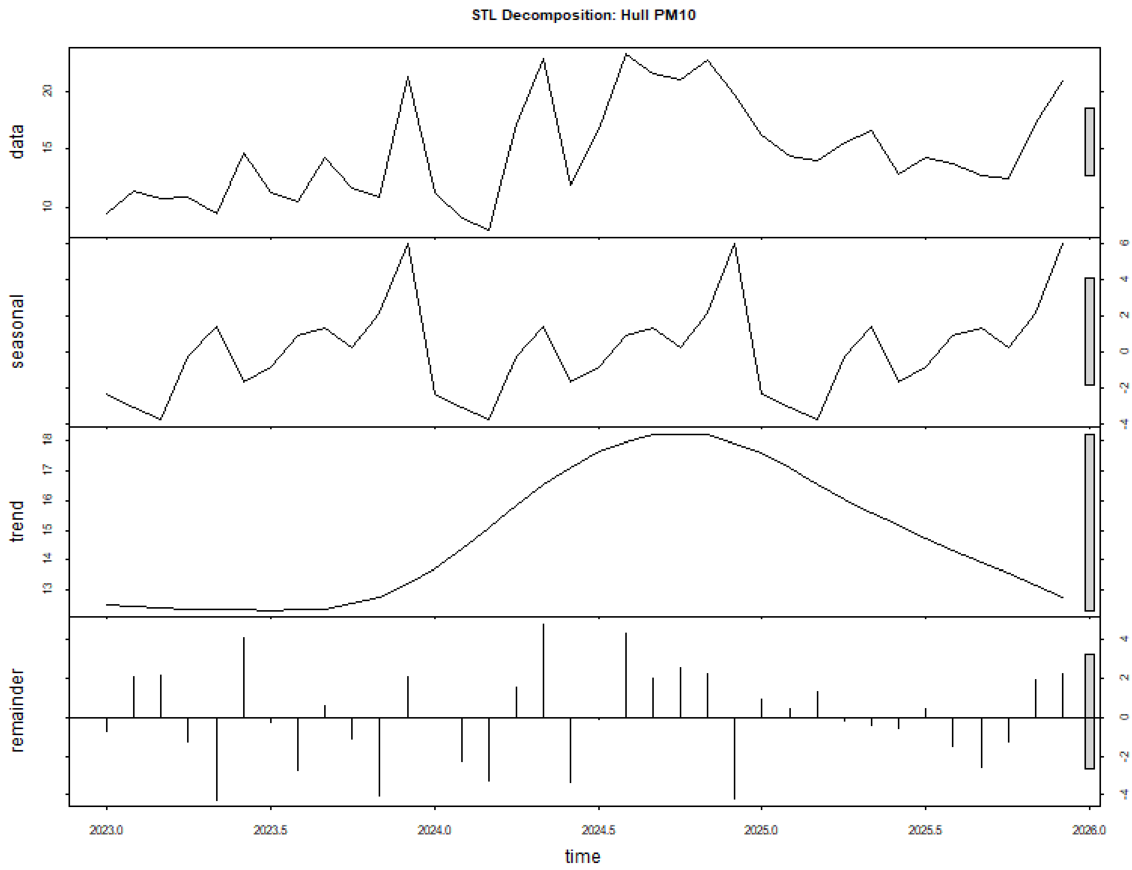


Figure 3.3 clarifies the temporal structure obscured in Figure 3.2. NO_2 concentrations decline from 2023 into late 2024, rise briefly in early 2025, and then decline again towards the end of the study period. $PM_{2.5}$ and PM_{10} show similar seasonal behaviour, with winter increases and summer declines. O_3 follows an opposing pattern, increasing during the warmer months and showing a gradual upward trend over time.

Compared with Figure 3.2, Figure 3.3 helps distinguish real trends from random daily fluctuations. The raw data are variable and erratic, especially in 2025, making it difficult to discern the direction of change. Smoothing provides a more precise representation of pollutant behaviour over time, confirming shared regional seasonal cycles while still emphasising differences in the emission profiles of Hull and Sheffield.

3.1.2 Seasonal Trend Decomposition using LOESS (STL)

Figure 3.4: STL Decomposition Hull PM_{10}



NO_2 and particulate matter demonstrate pronounced seasonal cycles, characterised by elevated concentrations in winter and lower levels in summer. Ozone shows peak concentrations during summer and a sustained upward trend, which aligns with reduced NO_2 driven ozone titration. The remaining components contribute minimally, indicating that seasonal and trend effects account for most of the observed variability rather than random fluctuations.

Figure 3.5: STL Decomposition Sheffield PM_{10}

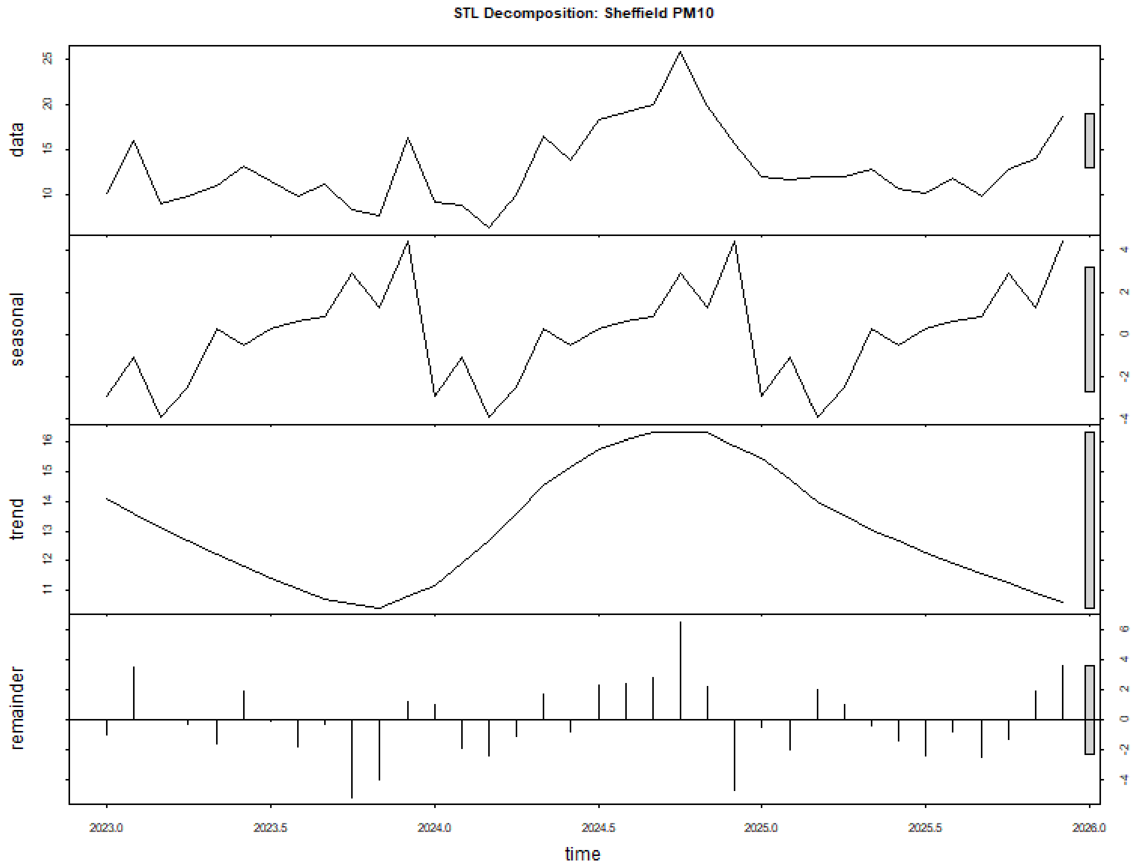


Figure 3.6: STL Decomposition Hull $PM_{2.5}$

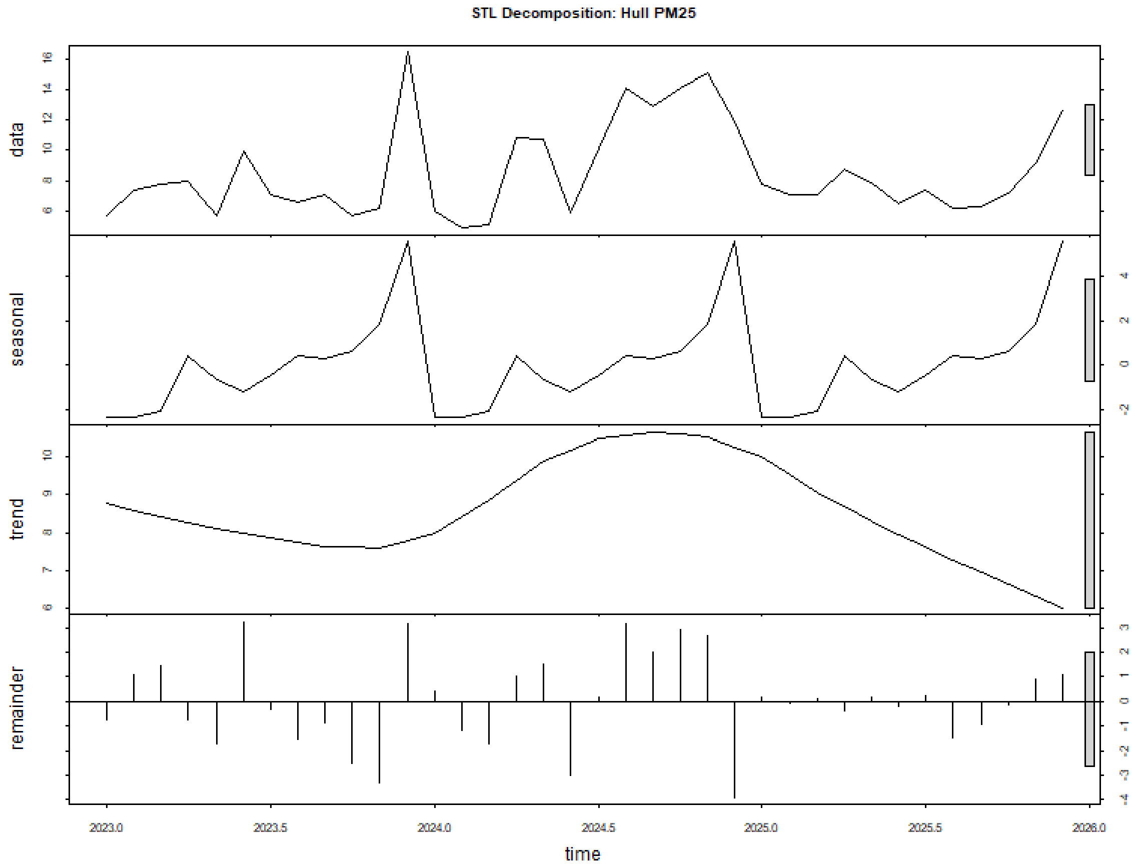


Figure 3.7: STL Decomposition Sheffield $PM_{2.5}$

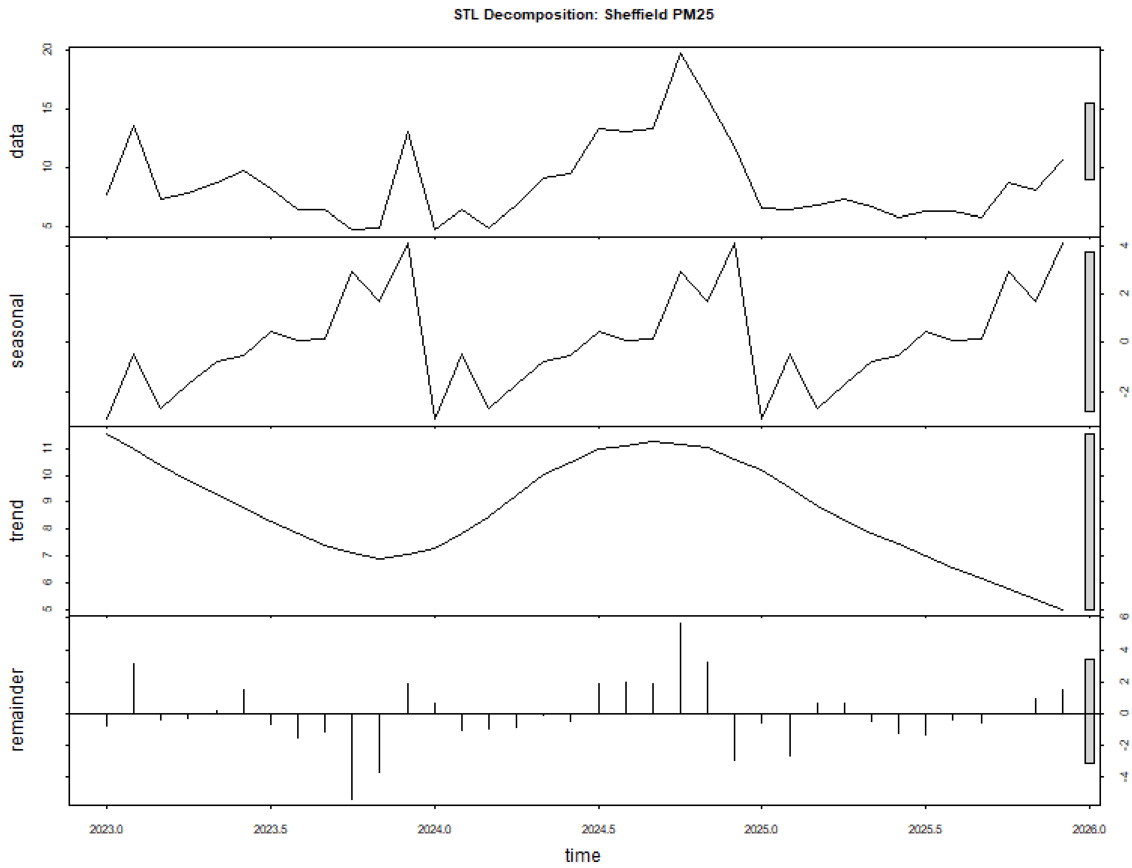


Figure 3.8: STL Decomposition Hull NO_2

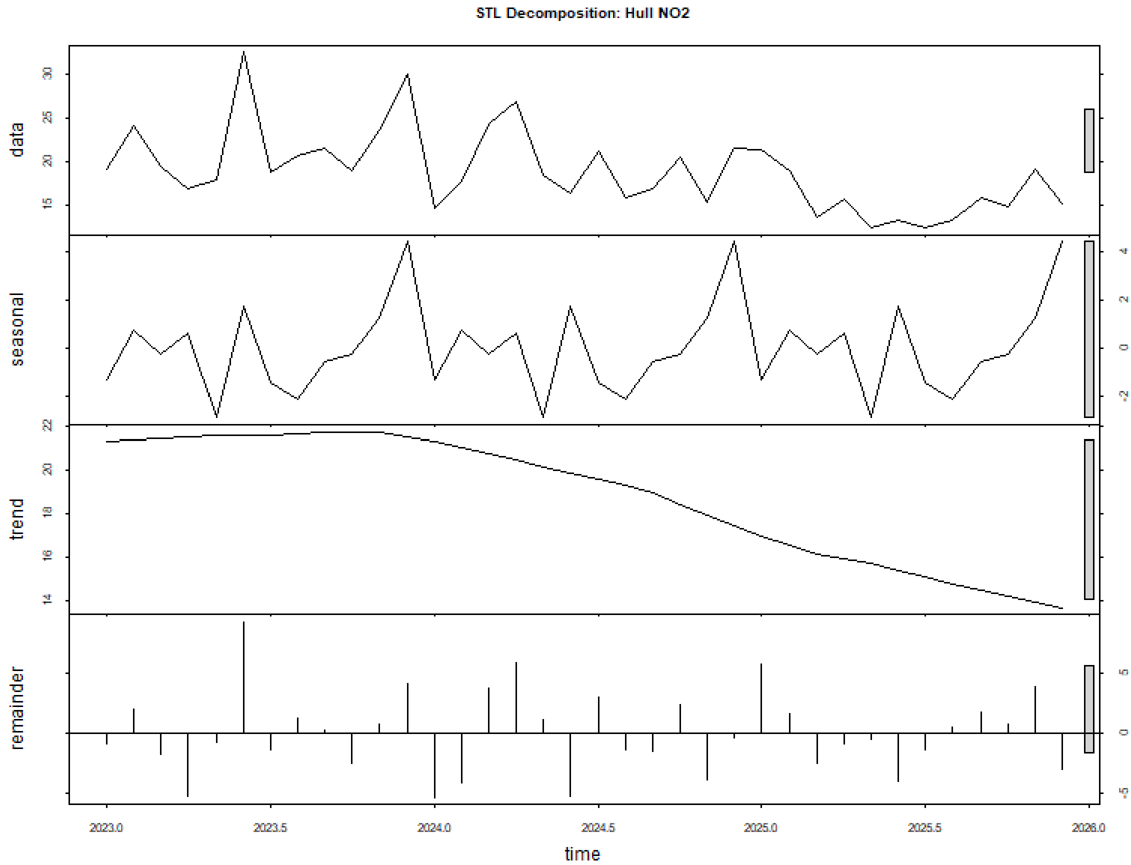


Figure 3.9: STL Decomposition Sheffield NO_2

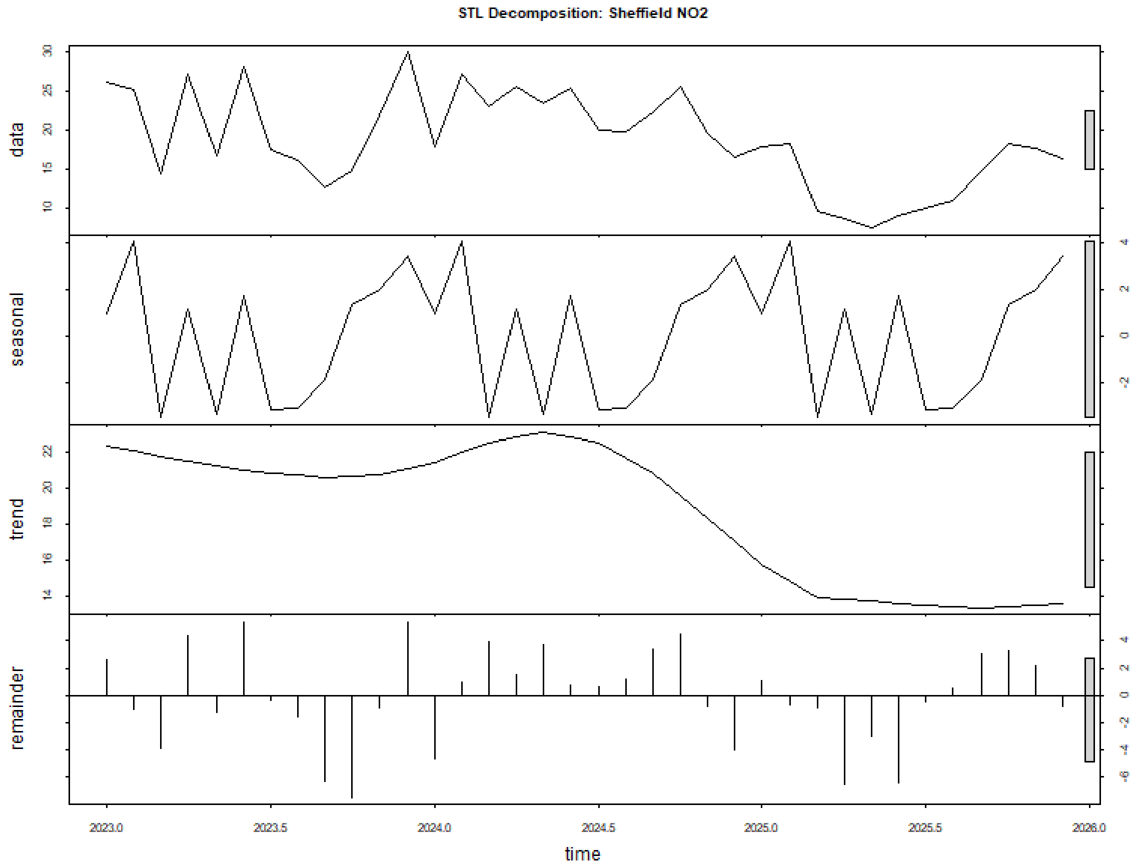


Figure 3.10: STL Decomposition Hull O_3

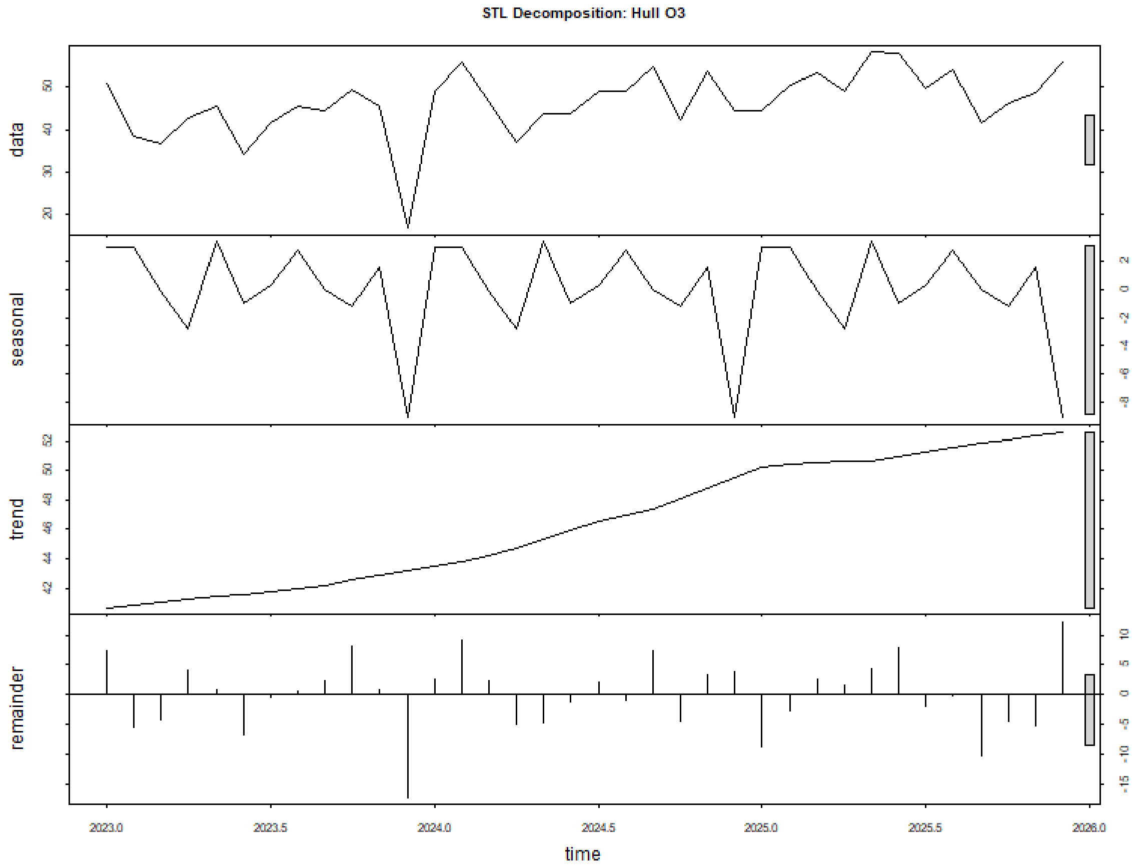
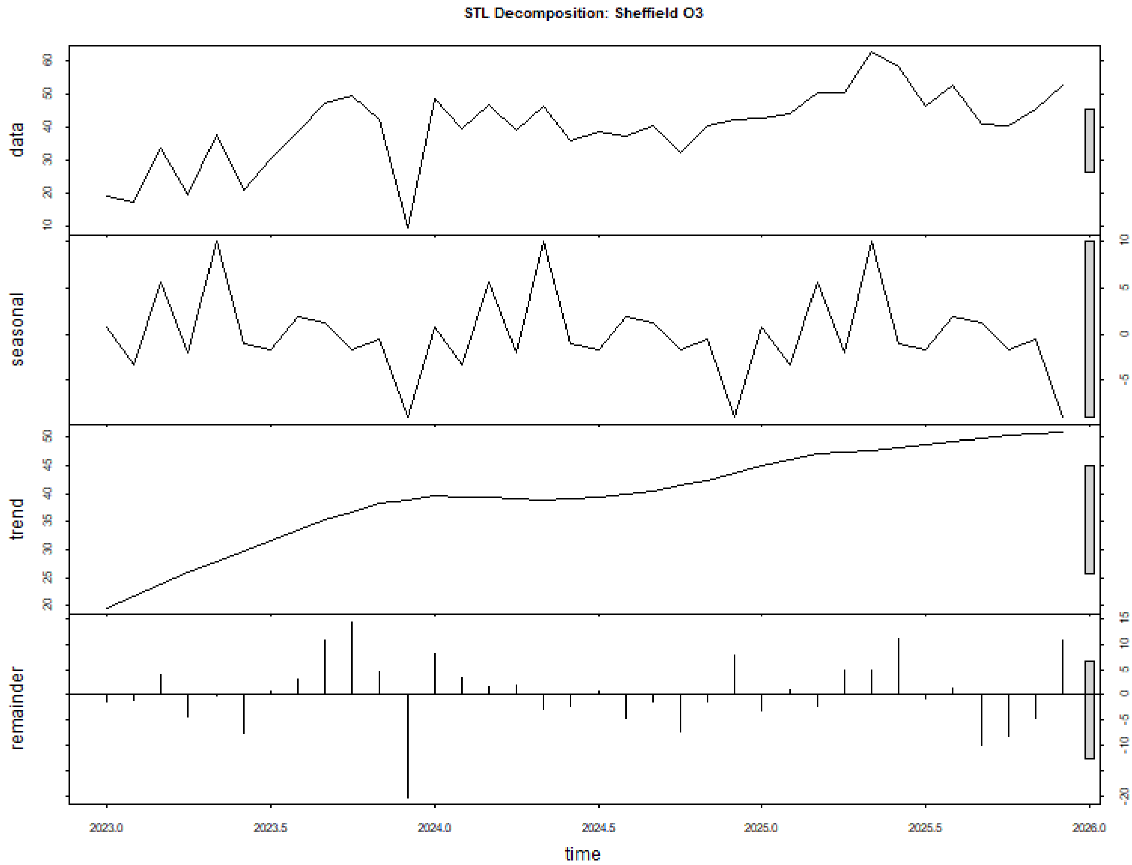
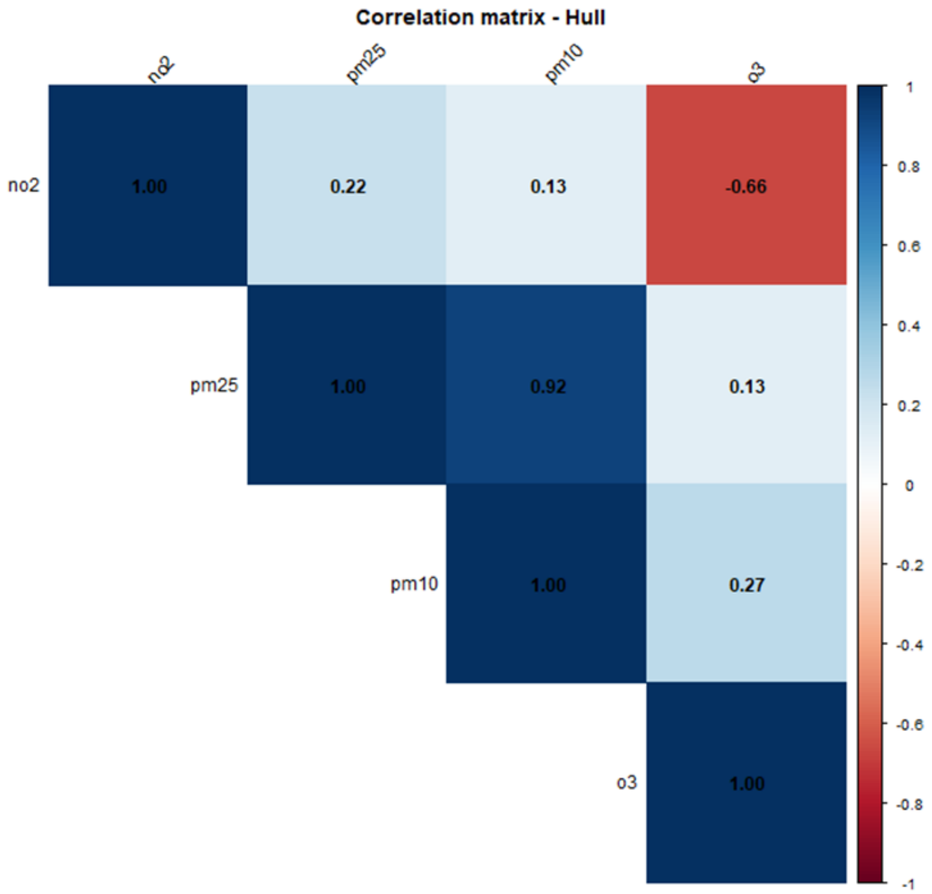


Figure 3.11: STL Decomposition Sheffield O_3



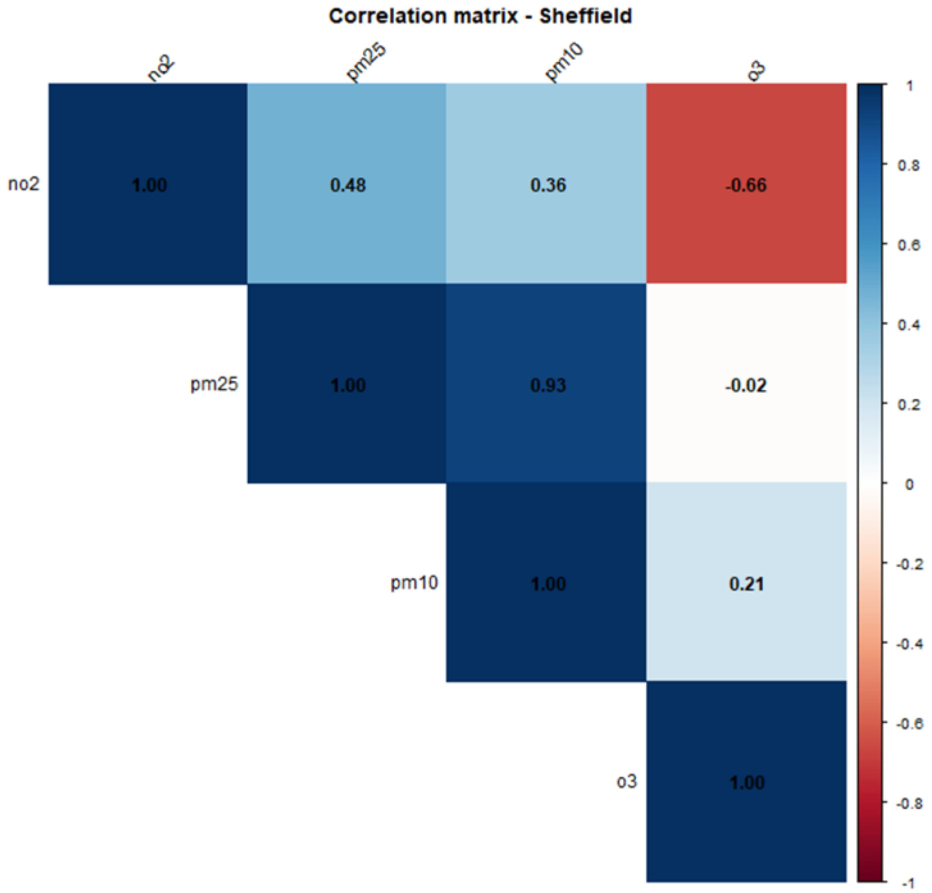
3.1.3 Correlation Matrix

Figure 3.12: Correlation Matrix Hull



Both cities show strong positive correlations between $PM_{2.5}$ and PM_{10}, suggesting shared sources or meteorological factors. In Sheffield, moderate positive correlations between NO_2 and particulates indicate traffic emissions. These correlations are weaker in Hull, likely due to more varied sources or greater coastal dispersion. Both cities also display a strong negative correlation between NO_2 and O_3, consistent with photochemical titration processes.

Figure 3.12: Correlation Matrix Sheffield



3.2 Discussion

LOESS and STL decomposition were essential for interpreting the data. These techniques separated trends from short-term noise and identified seasonal patterns. Correlation analysis supported existing chemical principles, especially the strong relationship between particulates and the inverse relationship between NO_2 and O_3.

3.2.1 Answering the Research Questions

3.2.1.1 How do key air pollutants vary between Sheffield and Hull from 2023 to 2025?

Overall, Sheffield consistently recorded higher NO_2 and $PM_{2.5}$ concentrations, while Hull generally exhibited higher PM_{10} and O_3 levels. These differences reflect contrasting emission profiles and environmental settings. Sheffield's elevated NO_2 and $PM_{2.5}$ are likely linked to higher traffic density and combustion sources, whereas Hull's higher PM_{10} and ozone appear influenced by industrial activity, maritime sources, and coastal atmospheric conditions. Although NO_2 declined in both cities over time, Sheffield remained consistently higher, whereas ozone gradually increased, with Hull maintaining higher levels throughout.

3.2.1.2 What daily, monthly, and seasonal patterns can be observed in pollutant levels across the two cities?

Daily pollutant concentrations display substantial short-term variability and occasional spikes in both cities, primarily influenced by meteorological conditions and transient events. At monthly and seasonal timescales, a more defined pattern is observed; concentrations of NO_2, $PM_{2.5}$, and PM_{10} reach their highest levels in winter and decrease during summer, reflecting increased emissions and reduced atmospheric dispersion in colder months. In contrast, ozone exhibits higher concentrations in spring and summer, attributable to intensified photochemical activity. STL decomposition shows that pronounced seasonal cycles account for much of the observed variability. Correlation analysis reveals consistent interrelationships among pollutants, particularly a strong negative association between NO_2 and O_3.

3.2.2 Relation to Existing Research

These results are consistent with existing UK air quality research. Sheffield's higher NO_2 and $PM_{2.5}$ levels reflect findings that traffic-dominated inland cities experience more combustion-related pollutants (Carslaw & Beevers, 2005; DEFRA, 2023).

In contrast, Hull's higher PM_{10} and ozone levels reflect patterns commonly reported in coastal and industrial cities, where enhanced atmospheric mixing and reduced NO_2 titration influence pollutant behaviour (Monks et al., 2015; Sicard et al., 2020).

The distinct seasonal patterns in both cities align with known atmospheric processes, with winter peaks in NO_2 and particulate matter, and spring-to-summer peaks in ozone driven by photochemical activity (WHO, 2021). The decline in NO_2 and rise in ozone further support evidence that changes in nitrogen oxide emissions can alter ozone chemistry. Overall, the findings reinforce existing literature and highlight the role of environmental context in shaping pollution patterns.

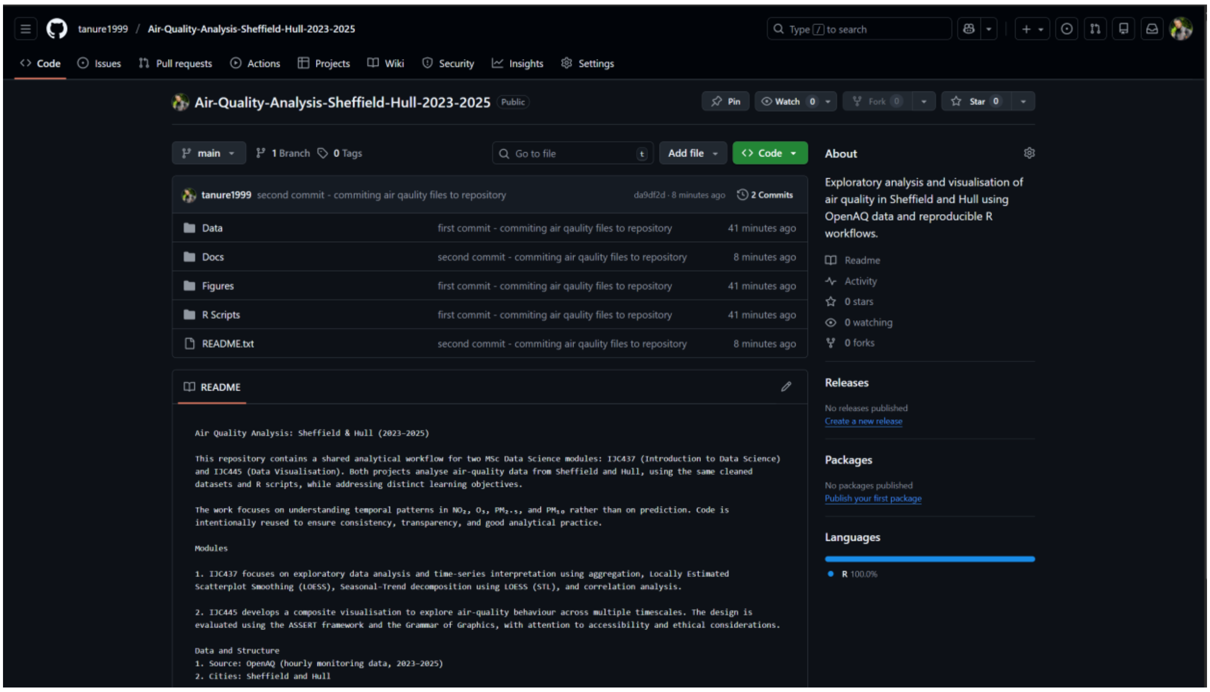
Chapter 4

GitHub

As part of the project workflow, all R scripts, datasets, and visualisation figures were uploaded to a GitHub repository. The repository was organised for easy navigation, with separate sections for the IJC437 and IJC445 project pages, and a code directory containing well-commented R scripts. Clear instructions for running the analysis were also included.

GitHub repository: https://github.com/tanure1999/Air-Quality-Analysis-Sheffield-Hull-2023-2025

Figure 4.1: GitHub Repo Page



Chapter 5

Conclusion

This study reveals clear seasonal patterns, distinct city-level pollution profiles, and consistent relationships among key pollutants. Sheffield consistently recorded higher NO_2 and $PM_{2.5}$ concentrations, reflecting traffic and combustion emissions, while Hull recorded higher PM_{10} and slightly elevated ozone levels, consistent with industrial activity and coastal dispersion.

Both cities exhibited strong seasonality, with winter peaks in NO_2 and particulate matter and summer peaks in ozone. Longer-term trends indicate declining NO_2 alongside a gradual rise in ozone, aligning with established atmospheric chemistry and changing emission policies.

The analysis was limited by uneven data coverage, reliance on a single monitoring station, and the potential masking of short-term pollution patterns through aggregation. Future work could integrate meteorological variables, include additional monitoring sites, and extend the analysis using predictive approaches.

Overall, the study demonstrates how combining statistical analysis with domain knowledge supports the interpretation of air quality data.

Reference

Carslaw, D. C., & Beevers, S. D. (2005). Estimations of road vehicle primary NO_2 exhaust emission fractions using monitoring data in London. *Atmospheric Environment, 39*(1), 167–177. https://doi.org/10.1016/j.atmosenv.2004.08.053

Cairo, A. (2016). *The truthful art: Data, charts, and maps for communication*. New Riders.

Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association, 79*(387), 531–554. https://doi.org/10.1126/science.229.4716.828

Department for Environment, Food & Rural Affairs (DEFRA). (2023). *Air quality statistics in the UK*. UK Government. https://www.gov.uk/government/statistics/air-quality-statistics

Few, S. (2012). *Show me the numbers: Designing tables and graphs to enlighten* (2nd ed.). Analytics Press.

Hull City Council. (2023). Air quality annual status report (ASR).

Monks, P. S., Archibald, A. T., Colette, A., Cooper, O., Coyle, M., Derwent, R., Fowler, D., Granier, C., Law, K. S., Mills, G. E., Stevenson, D. S., Tarasova, O., Thouret, V., von Schneidemesser, E., Sommariva, R., Wild, O., & Williams, M. L. (2015). Tropospheric ozone and its precursors from the urban to the global scale from air quality to short-lived climate forcer. *Atmospheric Chemistry and Physics, 15*(15), 8889–8973. https://doi.org/10.5194/acp-15-8889-2015

OpenAQ. (2024). OpenAQ platform: Open air quality data. https://openaq.org

Sheffield City Council. (2010). Air quality in Sheffield.

https://www.sheffield.gov.uk/pollution-nuisance/air-quality

Sicard, P., De Marco, A., Agathokleous, E., Feng, Z., Xu, X., Paoletti, E., & Calatayud, V. (2020). Amplified ozone pollution in cities during the COVID-19 lockdown. *Science of the Total Environment, 735*, 139542. https://doi.org/10.1016/j.scitotenv.2020.139542

Ware, C. (2013). *Information visualisation: Perception for design* (3rd ed.). Morgan Kaufmann.

World Health Organisation (WHO). (2021). *WHO global air quality guidelines: Particulate matter ($PM_{2.5}$ and PM_{10}), ozone, nitrogen dioxide, sulphur dioxide and carbon monoxide*. https://www.who.int/publications/i/item/9789240034228

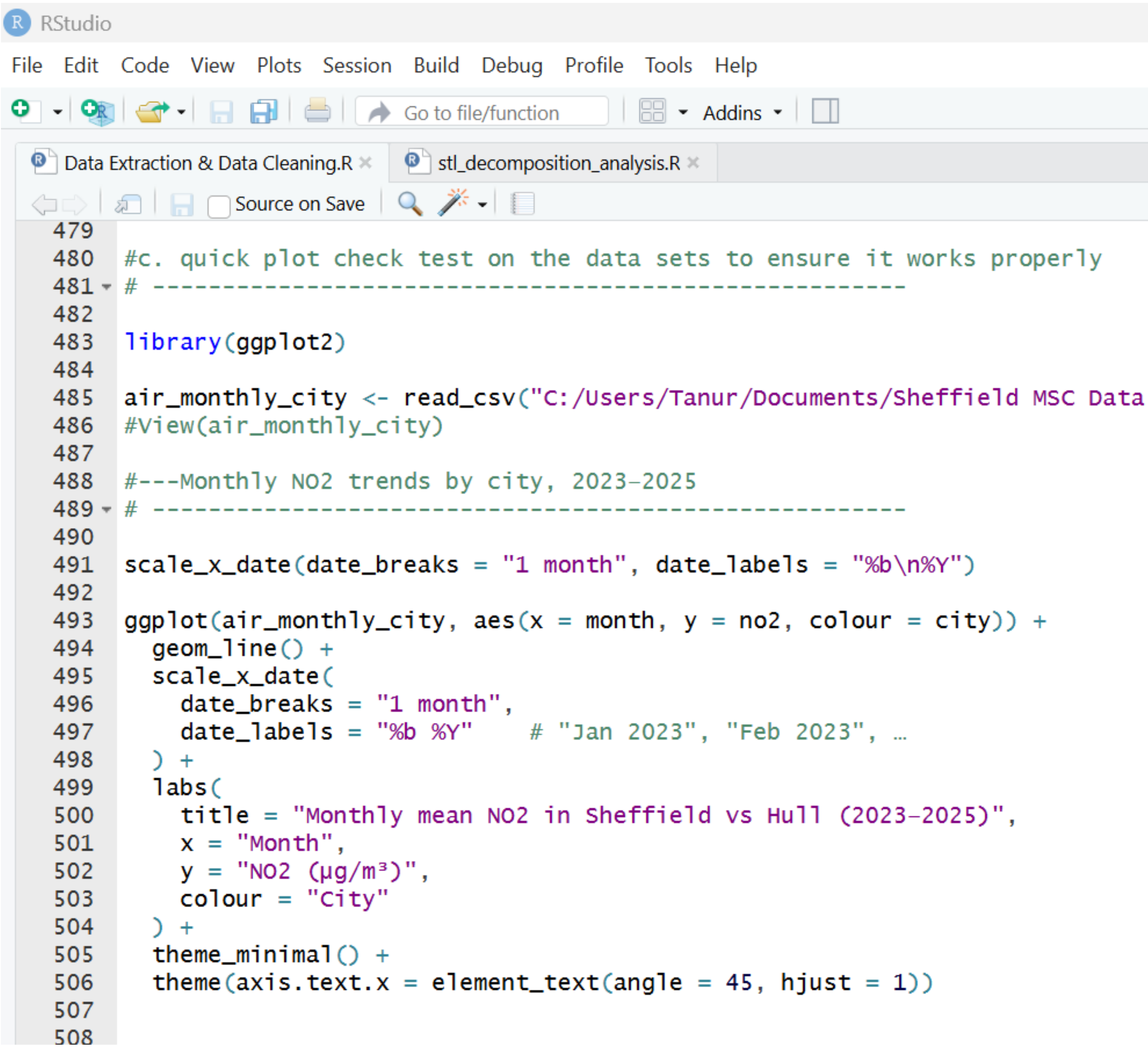
World Wide Web Consortium (W3C). (2018). *Web Content Accessibility Guidelines (WCAG) 2.1*. https://www.w3.org/TR/WCAG21/

Appendix

7.1: Install & Load Packages

```r
#DATA EXTRACTION
#Data Cleaning
#Data Transformation

# -----------------------------------------------------
#1. install packages (to run excel exl files)
# -----------------------------------------------------
install.packages("writeexl")
install.packages("openxlsx")

# -----------------------------------------------------
#2. load packages
# -----------------------------------------------------


library(tidyverse)
library(tidyr)
library(dplyr)
library(purrr)
library(writexl)
library(ggplot2)
library(lubridate)
library(readr)
#Run Library
library(tidyverse)
library(lubridate)
library(forecast)   # for STL + autoplot
library(corrplot)   # for nice correlation plots

```

7.2: Quick Plot Test After Data Cleaning




```r

#c. quick plot check test on the data sets to ensure it works properly
# ------------------------------------------------------

library(ggplot2)

air_monthly_city <- read_csv("C:/Users/Tanur/Documents/Sheffield MSC Data
#View(air_monthly_city)

#---Monthly NO2 trends by city, 2023-2025
# ------------------------------------------------------

scale_x_date(date_breaks = "1 month", date_labels = "%b\n%Y")

ggplot(air_monthly_city, aes(x = month, y = no2, colour = city)) +
  geom_line() +
  scale_x_date(
    date_breaks = "1 month",
    date_labels = "%b %Y"     # "Jan 2023", "Feb 2023", …
  ) +
  labs(
    title = "Monthly mean NO2 in Sheffield vs Hull (2023-2025)",
    x = "Month",
    y = "NO2 (µg/m³)",
    colour = "City"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))


```

7.3: CSV(not all) Files after Data Extraction & Cleaning

```r
# ------------------------------------------------------
# ------------------------------------------------------
# ------------------------------------------------------
#IMPORTANT NOTICE FOR NEXT STEP DATA ANALYSIS
# ------------------------------------------------------
#Load new csv files after manual cleaning with Excel
# ------------------------------------------------------
air_daily_city_2023_2024 <- read_csv("data sets/air_daily_city_2023-2024.csv") #mean average of 20th & 21st jan - dec 2023-2024
air_daily_city_2025 <- read_csv("data sets/air_daily_city_2025.csv") #mean average per day 1st to 31st, jan to dec 2025
air_monthly_city_2023_2024_2025 <- read_csv("data sets/air_monthly_city.csv") #mean average per month (jan - dec) 2023-2025
air_hourly_2023 <- read_csv("data sets/air_quality_hourly_2023.csv") #hourly parameters jan to dec 2023
air_hourly_2024 <- read_csv("data sets/air_quality_hourly_2024.csv") #hourly parameters jan to dec 2024
air_hourly_2025 <- read_csv("data sets/air_quality_hourly_2025.csv")#hourly parameters jan to dec 2025
# ------------------------------------------------------
# ------------------------------------------------------
# ------------------------------------------------------
```

FIG 7.4: Load Aggregated CSV Files

```
# Create folders for outputs
dir.create("C:/Users/Tanur/Documents/Sheffield MSC Data Science/INTRODUCTION TO DATA SCIENCE
dir.create("C:/Users/Tanur/Documents/Sheffield MSC Data Science/INTRODUCTION TO DATA SCIENCE
dir.create("C:/Users/Tanur/Documents/Sheffield MSC Data Science/INTRODUCTION TO DATA SCIENCE
dir.create("C:/Users/Tanur/Documents/Sheffield MSC Data Science/INTRODUCTION TO DATA SCIENCE

#-------------------------------------------------------------------------

#Section 2
#Load CSV Files
# ---- 1. LOAD DATA ----

# Change paths to where your CSVs are saved
daily <- read_csv("C:/Users/Tanur/Documents/Sheffield MSC Data Science/INTRODUCTION TO DATA S
monthly <- read_csv("C:/Users/Tanur/Documents/Sheffield MSC Data Science/INTRODUCTION TO DATA
seasonal <- read_csv("C:/Users/Tanur/Documents/Sheffield MSC Data Science/INTRODUCTION TO DAT

view(monthly)
# Quick checks
glimpse(daily)
glimpse(monthly)
glimpse(seasonal)

#
```

FIG 7.5: Standardising Date Columns

```
#Section 3
#Clean & Standardise Date Columns
#this is done as a double check
#---- 1. Daily: ISO dates to Date-----

daily <- daily %>%
  mutate(
    date = mdy(date),
    city = factor(city),
    location_name = factor(location_name),
    year = year(date)    # recompute to be safe
  )

glimpse(daily)
# Daily data: date is already "YYYY-MM-DD"


#---- 2. Monthly: month string "2023-01-01T00:00:00Z" to Date-----
# Monthly data: parse month string properly
monthly <- monthly %>%
  mutate(
    date = as.Date(month),
    city = factor(city),
    year = year(month)    # recompute to be safe
  )

glimpse(monthly)
# Monthly data: date is already "YYYY-MM-DD"
```

FIG 7.6: Seasonal Bar Plot

```r
#plot
p_seasonal_month_2025 <- monthly_seasonal_2025 %>%
  ggplot(aes(x = season, y = season_mean, fill = city)) +
  geom_col(position = "dodge") +
  facet_wrap(~ pollutant, scales = "free_y") +
  labs(
    title = "Seasonal average pollutant concentrations 2025 (from monthly data)",
    x = "Season",
    y = "Mean concentration (µg/m³)",
    fill = "City"
  ) +
  theme_minimal()

p_seasonal_month_2025

ggsave("C:/Users/Tanur/Documents/Sheffield MSC Data Science/INTRODUCTION TO DATA SCIENCE IJC 437/assessment/r studio/Assessment Intro to Data Science/plots/seasonal_means_by_city_2025.png",
       p_seasonal, width = 10, height = 6, dpi = 300)
```

FIG 7.7: Scatter Plot

```r
# Scatter plot only
p_seasonal_scatter <- ggplot(seasonal_long,
                             aes(x = date, y = value, colour = city)) +
  geom_point(alpha = 0.4, size = 0.7) +      # scatter plot
  facet_wrap(~ pollutant, scales = "free_y") +
  labs(
    title = "Daily Pollutant Concentrations (Scatter Plot)",
    x = "Date",
    y = "Concentration (µg/m³)",
    colour = "City"
  ) +
  theme_minimal()

p_seasonal_scatter

# Save the scatter plot
ggsave("C:/Users/Tanur/Documents/Sheffield MSC Data Science/INTRODUCTION TO DATA SCIENCE IJC 437/assessment/r studio/Assessment Intro to Data Science/plots/scatter_only.png",
       p_seasonal_scatter, width = 10, height = 6, dpi = 300)
```

FIG 7.8: LOESS Plot

```r
#---------------------------------------------
#Section 7
# Time-Series with LOESS (for EDA)

# ---- 1. DAILY TIME SERIES + LOESS ----

p_daily_loess <- ggplot(daily_long,
                        aes(x = date, y = value, colour = city)) +
  geom_point(alpha = 0.2, size = 0.4) +
  geom_smooth(method = "loess", se = FALSE) +
  facet_wrap(~ pollutant, scales = "free_y") +
  labs(
    title = "Daily pollutant concentrations (LOESS-smoothed)",
    x = "Date",
    y = "Concentration (µg/m³)",
    colour = "City"
  ) +
  theme_minimal()

p_daily_loess

ggsave("C:/Users/Tanur/Documents/Sheffield MSC Data Science/INTRODUCTION TO DATA SCIENCE IJC 437/assessment/r studio/Assessment Intro to Data Science/plots/daily_LOESS_sheffield_vs_hull.png",
       p_daily_loess, width = 10, height = 6, dpi = 300)
```

FIG 7.9: STL Decomposition using LOESS

```r
# ---- 1. STL SEASONAL DECOMPOSITION ----

run_stl_for_series <- function(df, city_name, pollutant_name) {

  sub <- df %>%
    filter(city == city_name,
           pollutant == pollutant_name) %>%
    arrange(month)

  if (nrow(sub) == 0) {
    warning("No data for ", city_name, " - ", pollutant_name)
    return(NULL)
  }

  # Determine start year & month from actual data
  start_year <- year(min(sub$month))
  start_month <- month(min(sub$month))

  ts_data <- ts(
    sub$monthly_mean,
    start = c(start_year, start_month),
    frequency = 12
  )

  stl_fit <- stl(ts_data, s.window = "periodic")
```

FIG 7.10: STL Decomposition using LOESS (ctd)

```r
  # Plot base STL output
  plot(
    stl_fit,
    main = paste("STL Decomposition:", city_name, toupper(pollutant_name))
  )

  invisible(stl_fit)
}


# ---- 2. Run STL for all pollutants & cities ----

cities <- levels(monthly_long$city)
pollutants <- levels(monthly_long$pollutant)

# All STL plots into one PDF
pdf("C:/Users/Tanur/Documents/Sheffield MSC Data Science/INTRODUCTION TO DATA SCIENCE IJC 437/assessment/r studio/Assessment Intro to Data Science/plots/stl/all_stl_plots.pdf", width = 8, height = 6)
for (c in cities) {
  for (p in pollutants) {
    message("Running STL for: ", c, " - ", p)
    run_stl_for_series(monthly_long, c, p)
  }
}
dev.off()

# Separate PNGs (Images files)
```

FIG 7.11: Correlation Matrix Plot

```r
# ---- 7. CORRELATION MATRIX ----

plot_city_cor_matrix <- function(df, city_name) {
  sub <- df %>%
    filter(city == city_name) %>%
    select(no2, pm25, pm10, o3)   # order as you like

  cor_mat <- cor(sub, use = "pairwise.complete.obs")
  print(cor_mat)

  corrplot(
    cor_mat,
    method = "color",
    type = "upper",
    addCoef.col = "black",
    tl.col = "black",
    tl.srt = 45,
    title = paste("Correlation matrix -", city_name),
    mar = c(0, 0, 2, 0)
  )

  invisible(cor_mat)
}
```

FIG 7.12: Correlation Matrix Plot (ctd)

```r
# Sheffield
png("C:/Users/Tanur/Documents/Sheffield MSC Data Science/INTRODUCTION TO DATA SCIENCE IJC 437/assessment/r studio/
    Assessment Intro to Data Science/plots/correlation/corr_sheffield.png", width = 800, height = 600)
cor_sheffield <- plot_city_cor_matrix(daily, "Sheffield")
dev.off()

# Hull
png("C:/Users/Tanur/Documents/Sheffield MSC Data Science/
    INTRODUCTION TO DATA SCIENCE IJC 437/assessment/r studio/Assessment Intro to Data Science/plots/correlation/corr_hull.png", width = 800, height = 600)
cor_hull <- plot_city_cor_matrix(daily, "Hull")
dev.off()

# Save numeric matrices
write_csv(as.data.frame(cor_sheffield), "C:/Users/Tanur/Documents/Sheffield MSC Data Science/INTRODUCTION TO DATA SCIENCE IJC 437/assessment/r studio/Assessment
write_csv(as.data.frame(cor_hull), "C:/Users/Tanur/Documents/Sheffield MSC Data Science/INTRODUCTION TO DATA SCIENCE IJC 437/assessment/r studio/Assessment Intro
```