WARWICK BUSINESS SCHOOL
THE UNIVERSITY OF WARWICK

Masters Programmes: Dissertation Cover Sheet

Student Number:	**5522715**
Degree Course:	**MSBA**
Dissertation Title:	**Identifying Air Pollution Hotspots and Socio-Economic Impacts in South Asia**
Module Code:	**IB93Y0**
Submission Deadline:	**29th August, 2024**
Date Submitted:	**29th August, 2024**
Word Count:	**10000 (excluding Acknowledgement, Index, List of Abbreviations and Appendix)**
Number of Pages:	**26 (excluding Acknowledgement, Index, List of Abbreviations and Appendix)**
Have you used Artificial Intelligence (AI) in any part of this assignment?	**No.**

Academic Integrity Declaration

We're part of an academic community at Warwick. Whether studying, teaching, or researching, we're all taking part in an expert conversation which must meet standards of academic integrity. When we all meet these standards, we can take pride in our own academic achievements, as individuals and as an academic community.

Academic integrity means committing to honesty in academic work, giving credit where we've used others' ideas and being proud of our own achievements.

In submitting my work, I confirm that:

- I have read the guidance on academic integrity provided in the Student Handbook and understand the University regulations in relation to Academic Integrity. I am aware of the potential consequences of Academic Misconduct.
- I declare that the work is all my own, except where I have stated otherwise.
- No substantial part(s) of the work submitted here has also been submitted by me in other credit bearing assessments courses of study (other than in certain cases of a resubmission of a piece of work), and I acknowledge that if this has been done this may lead to an appropriate sanction.
- Where a generative Artificial Intelligence such as ChatGPT has been used I confirm I have abided by both the University guidance and specific requirements as set out in the Student Handbook and the Assessment brief. I have clearly acknowledged the use of any generative Artificial Intelligence in my submission, my reasoning for using it and which generative AI (or AIs) I have used. Except where indicated the work is otherwise entirely my own.
- I understand that should this piece of work raise concerns requiring investigation in relation to any of points above, it is possible that other work I have submitted for assessment will be checked, even if marks (provisional or confirmed) have been published.
- Where a proof-reader, paid or unpaid was used, I confirm that the proof-reader was made aware of and has complied with the University's proofreading policy.

Upon electronic submission of your assessment you will be required to agree to the statements above

Acknowledgement

I would like to express my deepest gratitude to the University of Warwick, for providing me with the opportunity and resources to pursue my academic goals. The support and facilities offered by the university have been instrumental in the completion of this dissertation.

I am particularly grateful to the Warwick Business School for their unwavering support and encouragement throughout my studies. The faculty and staff have been incredibly helpful, offering guidance and assistance whenever needed.

A special thanks to my supervisor, Dr. Kathryn Hoad, whose expertise, patience, and insightful feedback have been invaluable. Her dedication and commitment to my academic progress have been a source of inspiration, and I am deeply appreciative of the time and effort she has invested in my work.

Lastly, I would like to thank my family and friends for their continuous support and encouragement, which have been crucial in helping me stay focused and motivated.

Index

i. Index of Figures and Tables

Index of Figures:

Index of Tables:

ii. Abstract

The issue of air pollution presents a substantial environmental and public health problem, particularly in rapidly developing regions such as South Asia. This study aims to identify air pollution hotspots in South Asia by evaluating the patterns in PM10, PM2.5, and NO2 concentrations, and examining their relationship with various socio-economic factors. The research utilises a clustering analysis approach to detect areas with significantly elevated pollution levels and explores the socio-economic landscape of these hotspots.

The Literature Review chapter highlights the severity of air pollution on both human health and the environment, emphasising the need for targeted interventions. This study expands upon current understanding by conducting a detailed analysis of pollution patterns across South Asia, pinpointing specific urban areas that require immediate attention. The findings emphasise the importance of addressing not just the environmental but also the socio-economic implications of air pollution.

The key findings suggest that cities in the Indo-Gangetic Plains are the main areas with high levels of air pollution in South Asia, with land consumption rate and population growth rate being the primary contributors to elevated pollution levels in these hotspots. The analyses also reveal some surprising results, like access to public transport not having minimal effect on air pollution levels. These insights are essential for policymakers and urban planners in devising strategies to mitigate the adverse effects of air pollution.

In conclusion, this study contributes to the understanding of air pollution dynamics in South Asia by identifying critical hotspots and the socio-economic scenario. The research establishes a foundation for developing targeted interventions that can improve air quality and public health in the region, while also addressing broader socio-economic challenges. This work underscores the need for comprehensive policies that integrate environmental and socio-economic considerations to effectively combat air pollution in rapidly urbanising areas.

iii. List of Abbreviations

ADHD	Attention Deficit Hyperactivity Disorder
ASD	Autism Spectrum Disorder
ASEAN	Association of Southeast Asian Nations
COPD	Chronic Obstructive Pulmonary Disease
COVID	Coronavirus Disease
EDA	Exploratory Data Analysis
EEAS	European External Action Service
GDP	Gross Domestic Product
GHG	Greenhouse Gas
IGP	Indo-Gangetic Plains
IQR	Interquartile Range
LCR	Land Consumption Rate
LCRPGR	Ratio of Land Consumption Rate to Population Growth Rate
MAR	Missing at Random
MCAR	Missing Completely at Random
MNAR	Missing Not at Random
MRIO	Multi-Regional Input–Output
NASA	National Aeronautics and Space Administration
NIEHS	National Institute of Environmental Health Sciences
NO_x	Oxides of Nitrogen
NO	Nitrogen Oxide
NO_2	Nitrogen Dioxide
O_3	Ozone
PGR	Population Growth Rate
PM10	Particulate Matters of size 10 µm
PM2.5	Particulate Matters of size 2.5 µm
PM30	Particulate Matters of size 30 µm
SAARC	South Asian Association for Regional Cooperation
SO_2	Sulphur Dioxide
UFPs	Ultrafine Particles
UN	United Nations
US	United States
US EPA	United States Environmental Protection Agency
VOCs	Volatile Organic Compounds
WHO	World Health Organisation

1. Chapter 1: Introduction

1.1. Background and Context

Air pollution has emerged as a crucial environmental and public health concern in the 21st century. This issue is particularly severe in rapidly developing regions such as South Asia, which includes per Kumar (2012) countries within and around the Indian Subcontinent, such as India, Pakistan, and Bangladesh. Urbanisation, industrialisation, and increasing vehicular emissions are major contributors to rising levels of harmful pollutants in the atmosphere, leading to significant health risks and environmental degradation (Rodrigues, Bhattacharya and Cabete, 2023).

Air pollution hotspots are areas where pollutants' concentrations are especially high making them the focal points of concern in the region. The identification and analysis of these hotspots are crucial for devising targeted interventions to mitigate the adverse impacts of pollution on human health and the environment. Understanding the spatial distribution of air pollution and its sources is essential for policymakers, urban planners, and public health officials as they develop strategies to address this escalating issue.

1.2. Rationale for the Study

This study aims to identify the air pollution hotspots and examine the pollution patterns and year-on-year trends. Additionally, with this study, we aim to understand the socio-economic scenarios and trends, and their relationship with the pollution levels in these hotspot cities as compared to those in other cities.

The study seeks to provide a detailed understanding of air pollution patterns across the South Asian region, with a focus on identifying areas that require urgent attention.

1.3. Research Questions and Objectives

The primary objective of this research is to identify pollution hotspots in South Asia by analysing year-on-year data of PM10, PM2.5, and NO_2 concentrations. The study will further explore the relationship between these pollution levels and urban population data, as well as other socio-economic parameters, to assess trends and impacts. The research will specifically address the following objectives:

- **Objective 1**: To identify cities in South Asia that are pollution hotspots, with a focus on their PM10, PM2.5, and NO_2 concentration levels.
- **Objective 2**: To analyse the trends in pollution levels over the past years across these cities.
- **Objective 3**: To understand the relationship between pollution levels and various socio-economic aspects.

These objectives are aligned with the **research question**: *How can clustering analysis detect pollution hotspots in South Asia, and what are the socioeconomic consequences of these hotspots?*

1.4. Significance of this Study

The importance of this study rests in identifying specific urban areas where pollution levels are critically high. This research offers a targeted approach to pollution mitigation. Policymakers can utilise these findings to allocate resources more efficiently, prioritise

intervention strategies, and develop tailored solutions that address the unique challenges of each hotspot.

Furthermore, the examination of the socio-economic impacts is crucial for designing comprehensive policies that not only reduce pollution levels but also enhance the quality of life for affected populations.

1.5. Structure of the Report

The dissertation is structured into five chapters, each addressing a specific aspect of the research.

- **Chapter 1: Introduction** - Presents an overview of the research background, objectives, significance, and structure of the dissertation report.
- **Chapter 2: Literature Review** - Offers a detailed review of the existing literature on air pollution, its sources, effects, and the specific situation in South Asia, with a focus on identifying the air pollution hotspots.
- **Chapter 3: Methodology** - Describes the data collection methods, and analytical techniques used in the study.
- **Chapter 4: Results and Discussions** - Presents the results of the analyses, identifying pollution hotspots and discussing the relationship with the socio-economic factors.
- **Chapter 5: Conclusion** - Summarises the key findings and outlines the recommendations for future research focused on air pollution in South Asia.

2. Chapter 2: Literature Review

The Literature Review establishes the foundation for this study by contextualising the air pollution crisis within both global and South Asian frameworks. It begins with an overview of global air pollution challenges, which serves as a foundation for understanding the severity of the situation in South Asia, a region significantly impacted by rapid industrialisation and urbanisation. The review then focuses on identifying major pollution sources in South Asia, such as industrial emissions and vehicular exhaust, highlighting the need for targeted interventions, which aligns with the study's goal of pinpointing pollution hotspots.

Additionally, the review delves into methodologies for identifying these hotspots and explores the socio-economic impacts of air pollution, directly linking to the study's objectives of understanding correlations between pollution levels and socio-economic factors. By synthesising existing research, the Literature Review emphasises the importance of localised, data-driven analyses, providing a strong foundation for the study's aim to address specific gaps in understanding air pollution patterns and their socio-economic consequences in South Asia.

2.1. Air Pollution and its sources

Pollution is the process that involves the introduction of harmful materials, known as pollutants, into the environment. Natural pollutants (like volcanic ash) and man-made pollutants (such as industrial waste), when persist and accumulate in the environment over time, pollute the air (European Environment Agency, 2023; National Geographic, 2024). Air pollution specifically refers to the presence of harmful substances in the atmosphere, which can be solid particles, liquid droplets, or gases. There are two major classifications of air pollution, namely, outdoor and indoor air pollution (WHO, 2014; World Health Organization, 2019; Dass, Srivastava and Chaudhary, 2021). Outdoor air pollution is specifically caused by fine particulate matter which causes various respiratory and cardiovascular health conditions. Some of the most common sources of outdoor air pollution include residential appliances used for cooking and heating, automobiles, power plants and other industries, agriculture and forest fires (World Health Organization, 2019; Dass, Srivastava and Chaudhary, 2021).

2.2. Effects of Air Pollution

Upon further examination of the concept of particulate matters, we understand, from the explanation provided by Flood-Garibay, Angulo-Molina and Méndez-Rojas (2023), that these particulate matters or PMs are particles of size 30 to 2.5 μm (PM30, PM10, and PM2.5 are some examples), while the ones with less than 0.1 μm size are called ultrafine particles or UFPs. Additionally, US EPA (2016) explains that particulate matters with a diameter of 2.5 μm or less are also called fine particles or PM2.5. These fine particles and ultrafine particles due to their small size and large surface areas easily adsorb and carry metals, microorganisms and other pollutants into the human body posing serious health risks, effecting the heart, lungs, brain, liver, spleen, kidneys, pancreas, gastrointestinal tract, joints, reproductive system etc, causing acute and chronic respiratory and cardiovascular diseases, including asthma, COPD and even lung cancer in worst cases (World Health Organization, 2019; Chen *et al.*, 2022; Zhang *et al.*, 2024). In addition to particulate matter, Manisalidis *et al.* (2020) listed several other pollutants like nitrogen oxide (NO), sulphur dioxide (SO_2), dioxins and volatile organic compounds (VOCs), among others, as major contributors to air pollution. These pollutants do not just have physical but mental impacts as well. They can impact the neurodevelopment health of infants and children, resulting in complications like autism spectrum disorder or ASD and attention deficit hyperactivity disorder or ADHD (Theron *et al.*, 2021). In severe cases, air

pollution even causes death. Ritchie and Roser (2024) highlighted air pollution as one of the top risk factors for death globally, particularly in low-income countries. NIEHS (2023) added that air pollution is responsible for more than 6.5 million deaths each year worldwide.

Particulate matter and other air pollutants clearly pose significant health risks, affecting multiple organs and contributing to serious conditions such as respiratory diseases, neurological disorders, and even death, highlighting the urgent need for effective air quality management and mitigation strategies globally.

2.3. Climate Change: A global concern of air pollution

Air pollution today causes a global emergency, as its consequences extend beyond the domain of human health. Fowler *et al.* (2020) assert that poor air quality also impacts crop loss, biodiversity, and climate change. Air pollution influences global climate change by altering radiative forcing, increasing surface temperature, and influencing air quality through the emissions of pollutants (M. Fiore *et al.*, 2012). These pollutants and their precursors affect a region's air quality and climate. This emphasises the importance of cutting the emissions of methane and ozone precursors to stop regional and global warming. Anenberg *et al.* (2010) further supporting this argument states that ozone on ground level along with fine particulate matter, i.e. PM2.5, have increased significantly as compared to the pre-industrial era, which contributes worldwide to premature deaths. Further investigation by Silva *et al.* (2013) found that anthropogenic PM2.5 causes more than 2.1 million premature deaths, while anthropogenic O_3 was associated with approximately 472,000 premature deaths per year globally. On an overall level, outdoor air pollution is the main cause of around 3.3 million premature deaths per year worldwide (Lelieveld *et al.*, 2015). Asian countries like India and China, which have more prevalence of residential energy emissions, are the most impacted with the highest number of such deaths. Lelieveld *et al.* (2015) further states that traffic and power generation are the largest contributors to deaths related to air pollution, specifically in urban areas.

In contrast to prevailing research, Jacob and Winner (2009) argue that climate change effects air quality, as it directly impacts weather patterns causing uneven and irregular rains, and heatwaves among other climatic conditions, causing worsened air quality and public health. In 2009, a study focused on the US showed that increased temperatures due to climate change result in photochemical production of O_3 resulting in differentiated rise in O_3 levels across regions (Weaver *et al.*, 2009). Based on aforementioned references, we can infer that air pollution and climate change are very strongly correlated. This relationship was also identified by West *et al.* (2013), who discussed the co-benefits of reduction of pollutants, and climate change mitigation by global greenhouse gas emissions mitigation. Mitigating GHG emissions results in reduced emission of co-emitted pollutants directly resulting in improved air quality. Mitigating climate change also indirectly results in better air quality. Both of these benefits indirectly contribute to improved human health.

The above arguments prompt us to consider of a solution to address the problem of air quality and consequent climate change, to improve human health. In 1997, the United Nations realised that climate change is a global emergency that requires international cooperation. In order to address this issue an international treaty, named the Kyoto Protocol, was signed in Kyoto, Japan (United Nations Climate Change, 2021). The main objective of the protocol was that the industrialised countries would reduce their greenhouse gas emissions by ~5% below the 1990 levels for a commitment period from 2008 to 2012. The Kyoto Protocol was a landmark step taken towards international climate change policy by setting legally binding targets in terms of reduction of emissions by industrialised countries. It was followed by the Doha Amendment, which was adopted as a second commitment period starting from 2013 till

2020. Not limiting to just the industrialised countries, the Doha Amendment was accepted by 147 parties. These parties committed themselves to reducing their greenhouse gas emission levels by at least 18% below 1990 levels during the said commitment period. The next big development in this noble cause was the Paris Accord. It sets a long-term goal of reducing GHG emissions, intending to hold global temperature rise to below 2°C above pre-industrial levels with the efforts to limit it to 1.5°C. The agreement signed is known as the Paris Agreement, and currently, 195 parties have joined it (United Nations, 2015). In one of the articles, Savaresi (2016) argues that, unlike previous treaties, the Paris Agreement symbolises a collective effort to fight global climate change, as it includes acceptance of all the countries and parties around the globe, while also addressing the needs of developing countries. Savaresi (2016) labels the Paris Accord as a holistic, flexible and inclusive approach to tackling pollution and climate change. However, some recent articles, such as Röser *et al.* (2020) and Mor and Ghimire (2022), highlight the lack of data collection and reporting infrastructure, and sufficient financial resources as some of the challenges in proper and effective implementation of the nationally determined contributions or the NDCs laid down under the Paris Agreement.

2.4. Worsening situation in the developing world

The chronological order of how the Kyoto Protocol started from just the industrialised countries and slowly got the attention of all the countries including developing ones, correlates with how Fowler *et al.* (2020) describe air pollution hotspots have shifted from Europe and North America to Asia, attributing these shifts to human activities interaction along with technological advancements, and related regulatory actions. Additionally, the statement from Friedrich (2017) that “Almost 300 million children around the world are exposed to toxic levels of outdoor air pollution, and those growing up in low- and middle-income countries are most at risk” explains how severe the situation has become over years in under-developed and/or developing countries. Talking about the severe situation in developing countries, while focusing on Africa, Jiying, Beraud and Xicang (2023) explained how air pollution has a big burden on a country’s economic situation, as it effects health and productivity, impacting the overall quality of life of the people. Some of the solutions to this situation, as suggested in various articles, are investing in and supporting clean and green energy, promoting eco-innovation, making stronger regulatory frameworks and implementing environmental taxes, and most importantly public awareness (Chien *et al.*, 2021; Jiying, Beraud and Xicang, 2023). Mohsin *et al.* (2021) further investigate how developing economies, specifically Asian countries, can benefit strongly by a simple transition from non-renewable to renewable, in terms of sources of energy that they rely on. Increasing the deployment of renewable resources has a positive correlation with not just environmental sustainability but also with economic growth, as it would create more jobs and more innovation. Clearly, based on this information under-developed and/or developing economies are the ones which are on the front line facing the worst consequences of air pollution be it climate change or socio-economic effects, and hence would be more beneficial to switch to green and sustainable methods.

But why are only the developing countries facing these severe consequences of air pollution? The simple reason is rapid growth and economic activities. Huang, Sadiq and Chien (2023) explained this in the context of ASEAN countries. The study highlights transportation, rapid urbanisation and economic growth as major factors contributing to air pollution. In one of their prior studies, Huang, Sadiq and Chien (2021) explored urbanisation as a factor causing carbon emissions, since urban areas usually have a large number of economic activities and relatively higher energy consumption. The study also highlights that in addition to the previously discussed methods sustainable urban planning and natural resource utilisation can help address air pollution, specifically in the urban areas. In a study focused on a Chinese city,

Zhan *et al.* (2023) also found that urbanisation can aggravate pollution levels as a result of limited dispersion and certain meteorological conditions. Another finding of this study was that rapid urbanisation generates large amounts of PM2.5, PM10, NO_X and O_3. Wang (2018) again brought up three major aspects of urbanisation that cause air pollution: industrial activities, vehicular emissions, and energy consumption, in the context of global health impacts of poor air quality. The study also emphasised that the densely populated cities are the most affected.

Collectively these studies demonstrate a comprehensive and overarching insight that the rapidly urbanising cities in developing/ growing economies are experiencing the wrath of air pollution and climate change the most. The primary contributors to the pollution are the ever-rising PM2.5, PM10, NO_X and O_3 levels due to industrial activities, automobile emissions and energy consumption in these cities.

2.5. The South Asian view

The next big question now arising is: which are the cities or regions that are most affected? South Asia is one such region which is at the highest risk. Several articles (Guttikunda, Goel and Pant, 2014; Mogno *et al.*, 2021; Jabbar *et al.*, 2022; Chatterjee *et al.*, 2023) indicate the alarming situation in the region and also identify PM2.5 as the biggest threat. Jabbar *et al.* (2022) further include PM10, NO_X and SO_2 to the list of pollutants that contribute significantly to the problems in the region. On the other hand, Mogno *et al.* (2021) discuss the seasonal variability of pollution levels, specifically in the Indo-Gangetic Plains. The study revealed that high levels of pollution in winter and relatively low levels during rainy seasons are caused by seasonal variability of PM2.5 and organic aerosols.

The situation in South Asia is clearly alarming, and adding to it Balakrishnan *et al.* (2019) found that air pollution is also reducing life expectancy in India. The study also highlighted regional disparity, attributing higher pollution levels to the states with higher urbanisation and industrialisation causing air quality to worsen. While reiterating PM2.5, PM10, NO_X and SO_2 as major pollutants Guttikunda, Goel and Pant (2014), also identified that the urbanisation and industrialisation generate construction dust and biomass burning which particularly in metropolitan areas results in further air deterioration. Mehmood *et al.* (2021) and Roy *et al.* (2023) identified similar root causes and impacts of air pollution in Pakistan and Bangladesh as well respectively. Since the region as a whole is facing the same challenges in addressing air pollution, cooperation should undoubtedly be a key component of their solution. In their conclusions, Jabbar *et al.* (2022) and Chatterjee *et al.* (2023) also coherently suggest that collaborated efforts on the regional level are required from all South Asian countries to deal with the rising air pollution levels.

World Health Organization (2022) although mentions particulate matter, carbon monoxide, ozone, nitrogen dioxide, and sulphur dioxide as major pollutants, World Health Organization (2024) focuses only on pollutants mainly generated from human activities which cause most of the urban pollution, which is the major concern in the region we are focusing on, i.e., South Asia. These pollutants as listed by World Health Organization (2024) are PM2.5, PM10, and NO_2.

2.6. Identifying the hotspots

After developing a good understanding of the regions of most concern and the major contributing factors, our next step is to identify the cities that are the hotspots of air pollution within the region. Several studies have been conducted and a number of papers have been published discussing ways of detecting these hotspots.

A hotspot is an area with high levels of pollutants concentrated causing serious health concerns. Zhang, Hannigan and Lv (2021) propose a two-step approach using mobile sensing data to identify these air pollution hotspots. The method involves analysing mobile sensing data to detect places/ areas with high pollution concentrations, these areas are labelled as air pollution hotspots. However, Kozáková *et al.* (2019) followed a slightly different approach. They deployed atmospheric back-trajectory analysis and spatial analysis to pinpoint these hotspots. In another study based in an Iranian city, Habibi *et al.* (2017) also used spatial analysis techniques to identify high concentrations of pollution due to area-wise variations, to understand pollution distribution within the city. Whilst these studies directly work on identifying the air pollution hotspots, Moran and Kanemoto (2016) draw an interdependency between global supply chains and air pollution hotspots, they utilised multi-regional input–output, i.e. MRIO model to understand economic activities and their environmental impacts on global supply chains. The MRIO model then links this supply chain data to emission inventories, that quantify air pollution levels, to identify regions where air quality is influenced by demands in some other part of the world.

In brief, Moran and Kanemoto (2016) and Kozáková *et al.* (2019) both tracking a trajectory to identify the pollution hotspots rather than directly pinpointing them the way Zhang, Hannigan and Lv (2021) did, shows that there can be more than one way of identifying the air pollution hotspots, with above discussed ones being just a few examples. The main objective behind identifying these hotspots is to be able to implement the recommendations as suggested by Chien *et al.* (2021), and Jiying, Beraud and Xicang (2023) more effectively and efficiently to mitigate the issue of air pollution from the region.

2.7. Summary

To summarise the literature review chapter we, based on our synthesis of all the above information, can say that air pollution is the introduction of harmful substances called pollutants into the atmosphere. These pollutants pose severe health risks by infiltrating our bodies and affecting vital organs like the heart, lungs, brain, and more. This can lead to a range of serious health conditions, such as respiratory and cardiovascular diseases, and even lung cancer in some cases. Sadly, air pollution causes millions of premature deaths each year, of which low-income countries account for the major proportion. It also has a detrimental effect on mental health, contributing to conditions like autism and ADHD in children.

However, air pollution impacts us beyond just health, it also stimulates climate change. Pollutants alter radiative forcing and increase surface temperatures, creating an atmosphere that further endangers our environment and health. This complex relationship means that reducing pollutants can simultaneously mitigate climate change and its associated health risks. Recognising this, international efforts like the Kyoto Protocol, the Doha Amendment, and the Paris Agreement have been drafted and signed by countries around the world to collaboratively curb air pollution by cutting greenhouse emissions and improving air quality. These agreements emphasise the importance of collective global action, with the Paris Agreement uniting 195 countries and other parties under a commitment to limit global temperature rise to below 2°C above the pre-industrial levels. However, there are various data, reporting and financial infrastructure challenges in the proper implementation of nationally determined contributions.

Developing countries, especially in South Asia, are particularly hard-hit by air pollution due to rapid urbanisation and industrialisation. These regions suffer from high levels of PM2.5, PM10, and NO_2, which worsen during winter due to seasonal variability. This pollution crisis leads to reduced life expectancy and significant health burdens in the region. To combat this, it's crucial to endorse and invest in clean energy, promote eco-innovation, strengthen regulatory

frameworks, and raise public awareness. Identifying pollution hotspots using data science can help target interventions more effectively, reducing health risks and environmental damage, in a collective and collaborative effort by these countries.

3. Chapter 3: Methodology

In this part of the report, we discuss our overall approach, especially the data cleaning and pre-processing, and the analyses involved to meet the research objectives, which, are, first, to identify the air pollution hotspots in South Asia, and, second, to understand the pollution trends and their relationship with different socio-economic aspects.

This chapter broadly covers 2 stages of analyses that help us fulfil our research objectives:

- **Hotspot Identification**: As discussed in section 2.6 of the Literature Review chapter, there can be more than one way to identify hotspots, for this study the method of segmenting the cities, using cluster analysis and then labelling the ones with the highest air pollution as hotspots, is used. Cluster analysis or clustering, as explained by Wierzchoń and Kłopotek (2018), is a simple method that puts two or more objects which are more similar to each other than others into a group called a cluster. This is critical in addressing the **research objective 1** (refer to section 1.3).
- **Exploratory Data Analysis on Socio-economic Data**: The second phase of analyses, i.e. the exploratory data analysis, is further divided into two parts each addressing one of the other two research objectives.
    - **Year-on-Year Trends analysis**: Plotting year-on-year trend lines is very useful in analysing the pollution level trends, i.e. the **2nd research objective** (refer to section 1.3).
    - **Relation with socio-economic factors**: Correlation analyses and regression analyses help in addressing the **3rd research objective** (refer to section 1.3) by determining the relationship between pollution and socio-economic factors.

  Exploratory data analysis or EDA, according to Komorowski *et al.* (2016), is essential to understand relationships between variables using statistics and data visualisation.

Additionally, we also discuss in this chapter how to plot these hotspots and identify the pollution patterns. First step in the analysis phase is to procure the data and clean it for the above-mentioned steps.

3.1. Defining the Region under Study

Before the data cleaning and consequent analyses, it is crucial to first understand which countries South Asia comprises. Based on different cultural and geographical definitions it can vary which countries would be included in South Asia, East Asia and Central Asia. One such definition is that of SAARC, i.e. South Asian Association for Regional Cooperation (EEAS, 2021). SAARC was established in 1985 with the aim of giving people of South Asia a better quality of life. According to EEAS (2021), countries that are part of SAARC are Afghanistan, Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan, and Sri Lanka. Additionally, according to the Fig. 3.1., as published by the World Bank on GDP growth rates and forecast of South Asian countries, it considers Maldives, Sri Lanka, Bangladesh, Bhutan, India, Nepal, and Pakistan as South Asia, excluding Afghanistan due to lack of available data (WorldBank, 2024).

	Country fiscal year	Real GDP growth at constant market prices (percent)				Revision to forecast from October 2023 (percentage point)	
Calendar year basis		**2022**	**2023(e)**	**2024(f)**	**2025(f)**	**2024(f)**	**2025(f)**
South Asia region (excluding Afghanistan)		**5.7**	**6.6**	**6.0**	**6.1**	**0.4**	**0.3**
Maldives	January to December	13.9	4.0	4.7	5.2	-0.5	-0.3
Sri Lanka	January to December	-7.3	-2.3	2.2	2.5	0.5	0.1
Fiscal year basis		**21/22**	**22/23(e)**	**23/24(f)**	**24/25(f)**	**23/24(f)**	**24/25(f)**
Bangladesh	July to June	7.1	5.8	5.6	5.7	0.0	-0.1
Bhutan	July to June	4.8	4.6	4.9	5.7	0.9	1.1
India	April to March	9.7	7.0	7.5	6.6	1.2	0.2
Nepal	mid-July to mid-July	5.6	1.9	3.3	4.6	-0.6	-0.4
Pakistan	July to June	6.2	-0.2	1.8	2.3	0.1	0.0

Sources: World Bank Macro Poverty Outlook and World Bank staff calculations.
Note: (e) = estimate; (f) = forecast. GDP measured in average 2010-19 prices and market exchange rates. Pakistan is reported at factor cost. National accounts statistics for Afghanistan are not available. To estimate forecasts for regional aggregates in the calendar year, fiscal year forecasts are converted to the calendar year by taking the average of two consecutive fiscal years for Bangladesh, Bhutan, Nepal, and Pakistan because quarterly GDP forecasts are not available.

Fig. 3.1. (WorldBank, 2024) *GDP Growth Rates and Forecasts.*

So, based on these articles, this dissertation will define the South Asian region as comprising of eight countries, namely, Afghanistan, Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan, and Sri Lanka.

3.2. Data Collection

Now that we know which is our region of focus for this study as well as what our objectives are (refer to Section 1.3. Research Questions and Objectives), our next step is to decide what type of data to use and subsequently procure it. As the source(s) of data should be highly reliable, extensive and exhaustive for more reliability, robustness and validity of the data, results and consequent insights, for this study, there are two types of data that would serve the purpose.

3.2.1. Pollution Data

While there might be multiple online open sources to use data from, it is essential that the dataset is exhaustive enough to have information about the three most prevalent pollutants in South Asia which, as discussed in section 2.5 of the Literature Review chapter, are PM2.5, PM10 and NO_2. WHO maintains a robust and comprehensive database covering the entire globe, broken down at the city level, giving a more granular view of the region.

Therefore, WHO's 'Air quality database 2022 (V5)' dataset (https://www.who.int/publications/m/item/who-ambient-air-quality-database-(update-jan-2024); refer to Appendix I) is the most appropriate dataset for the purpose of hotspot identification. For this report, only a subset of this entire dataset is used. All the South Asian cities are filtered, and a new smaller dataset is created that can be used for the analysis. This data is used for hotspot identification using clustering.

The Excel format data file contains 3 tabs. Out of these tabs, the 'Update 2024 (V6.1)' tab contains the raw data, while the tabs named 'Metadata' and 'Readme' contain information about the dataset and variables present in the 'Update 2024 (V6.1)' tab.

Out of all the variables, only the below listed ones are used for the further analysis, since other variables are irrelevant (refer to Appendix II for details on rationale for removal of such variables).

Variable Name	Description	Data type
country_name	Country Name	Nominal Data (Text)
city	City Name	Nominal Data (Text)
year	Year of the annual mean concentration	Discrete Data
pm10_concentration	Annual mean concentration of PM10	Continuous Data
pm25_concentration	Annual mean concentration of PM2.5	Continuous Data
no2_concentration	Annual mean concentration of NO_2	Continuous Data
population	Population of the city	Continuous Data
latitude	Latitude	Continuous Data
longitude	Longitude	Continuous Data

Table 3.1. Summary of the variables in the 'Pollution' dataset.

Below is the logical schema of the hence obtained 'Pollution' dataset:

pollution (*country_name*, *city*, *year*, pm10_concentration, pm25_concentration, no2_concentration, population, latitude, longitude)

3.2.2. Socio-Economic Data

For the second phase of analyses, we need population data of all the countries and cities within the region. Although each government publishes its population census data, which are open and free to download and use, on one of its central statistics and data repository websites, it's better to procure data from one common source for consistency, as different data sources can result in conflict in terms of format, semantics, and values (Shi *et al.*, 2019). Therefore, keeping in mind the consistency in data, UN-Habitat's global urban database was referred for the population data. Specifically 3 datasets with South Asian coverage were chosen:

I. "Open & Green Spaces" data (https://data.unhabitat.org/pages/open-spaces-and-green-areas; refer to Appendix I)
   - Contains 'Average share of the built-up area of cities that is open space for public use for all (%)' and 'Average share of urban population with convenient access to open public spaces (%)' along with geographic information like country and city name

II. "Urban transport" data (https://data.unhabitat.org/pages/urban-transport; refer to Appendix I)
   - Contains 'Share of urban population with convenient access to public transport (%)' and country and city name

III. "Growth of urban areas" data (https://data.unhabitat.org/pages/spatial-growth-of-cities-and-urban-areas; refer to Appendix I)
   - Contains multi-temporal data on the Land consumption rate, Population growth rate, ratio of land consumption rate to population growth rate, and built-up area per capita along with country and city name

These 3 datasets are particularly suitable for the study, as the 'Literature Review' chapter indicates that urbanisation and transportation are the biggest contributors to air pollution in the South Asian region.

However, we first have to remove the unwanted variables. Following the similar approach as above, here is a list of variables from the 'Open & Green Spaces' data that are used for the analyses, while the rationale of removal of the other variables can be found in Appendix II.

Variable Name	Data type
Country or Territory Name	Nominal Data (Text)
City Name	Nominal Data (Text)
Average share of the built-up area of cities that is open space for public use for all (%) [a]	Continuous Data
Average share of urban population with convenient access to open public spaces (%) [b]	Continuous Data
Data Reference Year	Discrete Data

Table 3.2. Summary of the variables in the 'Open & Green Spaces' dataset.

Here is the list of variables that are used for the analyses from the 'Urban transport' dataset. Deprioritised variables are listed in the Appendix II along with their rationale for removal.

Variable Name	Data type
Country or Territory Name	Nominal Data (Text)
City Name	Nominal Data (Text)
Share of urban population with convenient access to public transport (%)	Continuous Data
Data Reference Year	Discrete Data

Table 3.3. Summary of the variables in the 'Urban transport' dataset.

Similarly, the below table shows the list of variables from the 'Growth of urban areas' datasets that are used for the analyses in this report (refer to Appendix II for rationale of removal of the remaining variables that are removed).

Variable Name	Data type
Country or Territory Name	Nominal Data (Text)
City Name	Nominal Data (Text)
Data Year 1	Discrete Data
Data Year 2	Discrete Data
Data Year 3	Discrete Data
Land consumption rate (LCR) Year 1 to Year 2 (%)	Continuous Data
Land consumption rate (LCR) Year 2 to Year 3 (%)	Continuous Data
Population Growth Rate (PGR) Year 1 to Year 2 (%)	Continuous Data
Population Growth Rate (PGR) Year 2 to Year 3 (%)	Continuous Data
Ratio of Land Consumption Rate to Population Growth Rate (LCRPGR) Year 1 to Year 2 (Ratio)	Continuous Data
Ratio of Land Consumption Rate to Population Growth Rate (LCRPGR) Year 2 to Year 3 (Ratio)	Continuous Data
Built Up area Per Capita Year 1 (m2 per person)	Continuous Data
Built Up area Per Capita Year 2 (m2 per person)	Continuous Data
Built Up area Per Capita Year 3 (m2 per person)	Continuous Data

Table 3.4. Summary of the variables in 'Growth of urban areas dataset.

The logical schemas of the hence obtained socio-economic datasets:

I. Open & Green Spaces

Open_and_green_spaces (*Country or Territory Name*, *City Name*, Average share of the built-up area of cities that is open space for public use for all (%) [a], Average share of urban population with convenient access to open public spaces (%) [b], Data Reference Year)

II. Urban transport

Urban_transport (*Country or Territory Name*, *City Name*, Share of urban population with convenient access to public transport (%), Data Reference Year)

III. Growth of urban areas

Growth_of_urban_areas (*Country or Territory Name*, *City Name*, Data Year 1, Data Year 2, Data Year 3, Land consumption rate (LCR) Year 1 to Year 2 (%), Land consumption rate (LCR) Year 2 to Year 3 (%), Population Growth Rate (PGR) Year 1 to Year 2 (%), Population Growth Rate (PGR) Year 2 to Year 3 (%), Ratio of Land Consumption Rate to Population Growth Rate (LCRPGR) Year 1 to Year 2 (Ratio), Ratio of Land Consumption Rate to Population Growth Rate (LCRPGR) Year 2 to Year 3 (Ratio), Built Up area Per Capita Year 1 (m2 per person), Built Up area Per Capita Year 2 (m2 per person), Built Up area Per Capita Year 3 (m2 per person))

3.3. Data Cleaning and Pre-processing

The 3 population/ socio-economic datasets along with the pollution data are uploaded on the Rstudio project for cleaning and pre-processing of the data before performing the analyses. Upon uploading the data, the below-listed steps are followed:

Step 1. *Filter out the non-South Asian countries.*

For this study, we only need to consider eight South Asian countries, as discussed in section 3.1. Therefore, we filter and keep only these eight countries for our analyses.

Step 2. *Filter out the unnecessary variables.*

After filtering out the countries, we select only those variables that are required for the analyses, as mentioned in the logical schemas formed in section 3.2 of the Methodology chapter.

Step 3. *Identifying and subsequently fixing the missing values.*

Several articles, such as Rubin (1996) and Enders (2022), discussed different ways of dealing with missing values with imputation being one of them. They also highlighted the importance of fixing missing values for accuracy and efficiency of the analyses, and to mitigate the impact on the findings. Whilst a number of publications, such as Salgado *et al.* (2016) and Kwak and Kim (2017), discuss 3 types of missing data, i.e. (i) MCAR or Missing Completely at Random, (ii) MAR or Missing at Random, and (iii) MNAR or Missing Not at Random, they also highlight deletion and imputation as most prevalent methods to deal with the missing data. Little and Rubin (2019) further discussed various methods of imputation and the importance of choosing the appropriate method for accurate and reliable statistical analysis. For our study, we need to carefully select and implement the most appropriate imputation and/ or deletion method(s)

to ensure the maximum possible data for more robust clustering results. The below table shows the proportions of missing values in each dataset.

Dataset Name	Proportion of missing values
Urban transport	0.00%
Open & Green Spaces	0.89%
Growth of urban areas	3.05%
Pollution	14.73%

Table 3.5. Proportions of missing values in each dataset.

The 'Urban transport' dataset is the only dataset which has no missing values. The 'Open & Green Spaces' dataset although has very few missing values, all of which are found only in the 'Average share of the built-up area of cities that is open space for public use for all (%) [a]' column. Upon filtering the data for the records with missing values, it is evident that there is still a diverse range of countries and cities present, therefore, it is deduced that the values are missing completely at random (MCAR). Therefore as suggested by Salgado *et al.* (2016), deletion method should be used to deal with these missing values. However, since the records with these missing values can be skipped while doing analyses on this variable, we keep these records rather than removing them altogether.

For the 'Growth of urban areas' dataset, all the missing values are in the 'Ratio of Land Consumption Rate to Population Growth Rate (LCRPGR) Year 2 to Year 3 (Ratio)' variable, which is basically a ratio of 'Land consumption rate (LCR) Year 2 to Year 3 (%)' and 'Population Growth Rate (PGR) Year 2 to Year 3 (%)'. Therefore, we can simply calculate all the missing values for this dataset. However, since the proportion of missing values in the 'Pollution' dataset is high, it is very important to impute its values for a robust clustering analysis. A random-forest based method, named 'missForest', is considered very flexible across various missing data mechanisms and has been shown to outperform other imputation strategies according to Stekhoven and Bühlmann (2012). Therefore, 'missForest' is a great method to deal with missing values in the 'Pollution' dataset.

Step 4. *Outlier detection.*

Once data is ensured to be complete, the next most essential data pre-processing step is the detection of the outliers. This step is very critical to identify extreme values, to ensure that they do not impact the results (Tiwari *et al.*, 2007). However, the outlier capping techniques are applicable only to the numerical data points which are not percentages, as the restricting intervals, i.e. 0-100, in the case of percentages, automatically serve as capping to such variables (Kieschnick and McCullough, 2003). Therefore, among the socio-economic datasets, outlier capping could only be applied on 'Built Up area Per Capita Year 1 (m2 per person)', 'Built Up area Per Capita Year 2 (m2 per person)', and 'Built Up area Per Capita Year 3 (m2 per person)' variables of the 'Growth of urban areas' dataset. Additionally, outliers are identified in 'Pollution' dataset as well since it is crucial for cluster analysis (Mahalanobis, 2018).

Step 4.1. Outlier detection in the 'Growth of urban areas' dataset.
While there are various methods of outlier removal and/ or capping, Rousseeuw and Hubert (2011) argued one such method, referring to Tukey's boxplot, and explained that the points lying outside the *fence* are labelled as outliers. Páez and Boisjoly (2022) also referred to these boxplots and explained that the box of a boxplot covers 50% of the records, while the stems extend to 1.5 times the interquartile range of the data, and any point outside these is named as an outlier. Once the potential outliers from the 'Growth of urban areas' dataset are identified using the IQR method, we check if capping the outliers can significantly change the statistical

measures like mean and median. Below is a summary of these statistical measures for both with and without outlier capping datasets.

Dataset	Mean 'Built Up area Year 1'	Mean 'Built Up area Year 2'	Mean 'Built Up area Year 3'
With Capping	64.47	68.50	73.30
Without Capping	44.27	47.29	51.69

Table 3.6. Mean values of three variables of 'Growth of urban areas' with and without capping.

Dataset	Median 'Built Up area Year 1'	Median 'Built Up area Year 2'	Median 'Built Up area Year 3'
With Capping	40.24	41.75	46.75
Without Capping	37.12	35.80	41.52

Table 3.7. Median values of three variables of 'Growth of urban areas' with and without capping.

Since the mean and median values are different in the 'With Capping' and 'Without Capping' datasets, we go ahead and check if they are statistically significant. The below table shows the t, df, and p-values from Welch's t-test on each of the above three variables.

Dataset	t	df	p-value
Built Up area Year 1	2.8888	133.95	0.004512
Built Up area Year 2	2.8514	137.41	0.005025
Built Up area Year 3	2.7865	142.6	0.006054

Table 3.8. Welch two sample t-test results.

Clearly, based on the Welch two sample t-test, the two datasets are significantly different, since the p-values are always less than 0.007. Therefore, we can call these 13 records as outliers from the 'Growth of urban areas' dataset. However, upon carefully checking, we notice that the 'Built Up area' values are consistently high for all the potential outlier records across the years. Therefore, these are correct and legitimate values which should not be removed from the data (refer to Appendix III for detailed table of these outliers).

Step 4.2. Outlier detection in the 'Pollution' dataset.
For the 'Pollution' dataset, we use the Mahalanobis distance method to identify the outliers, following the research, such as Dashdondov and Kim (2020, 2023), have done in past. They also calculate the p-values of these Mahalabonis distances and flag records with p-value less than 0.001, as the outliers. Below is the formula used by Dashdondov and Kim (2020, 2023) for Mahalanobis distance:

$$D^2 = (x - m)^T C^{-1} (x - m)$$

where, D is the Mahalanobis distance,
x is row in the dataset,
m is mean of each column, and
C is the covariance matrix of the independent variables.

Mahalanobis (2018) also demonstrated that Mahalanobis distance identifies outliers in multivariate data. It emphasised that this method enhances the ability to cluster data points accurately. Therefore, Mahalanobis distance technique is suitable to apply on the 'Pollution' dataset before going forward with the clustering analysis.

Following the Mahalanobis distance method, a total of 29 records were removed from the 'Pollution' dataset, resulting in a dataset with 1383 records for the clustering analysis in the next step.

3.4. Hotspot Identification

3.4.1. Creating Clusters

As discussed at the beginning of this chapter, clustering analysis is done to identify the air pollution hotspots. Clustering can be done by various methods, such as partitional, hierarchical, and density-based methods, as explained in various articles, such as Rokach and Maimon (2005), Xu and Wunsch (2005), and Aggarwal and Reddy (2014).

The k-means method is a feasible and appropriate clustering approach to group the cities based on pollution patterns as several studies, including Ahmad, Arshad and Sarlan (2022), have been based on the k-means clustering method with the aim to identify patterns and thereby form group or segments. Talking further about k-means clustering algorithms Ahmed, Seraj and Islam (2020) highlighted that the k-means technique has a diverse applicability, including but not limited to image processing and segmentation.

Since the k-means method requires a number as an input for the number of clusters to be formed, Kassambara and Mundt (2016) used the 'Gap statistic' to determine the optimal number to input. Murtagh and Contreras (2017) also endorsed hybrid models combining non-hierarchical techniques like k-means with hierarchical methods to do clustering. Below is the gap statistic plot built on the 'Pollution' data:

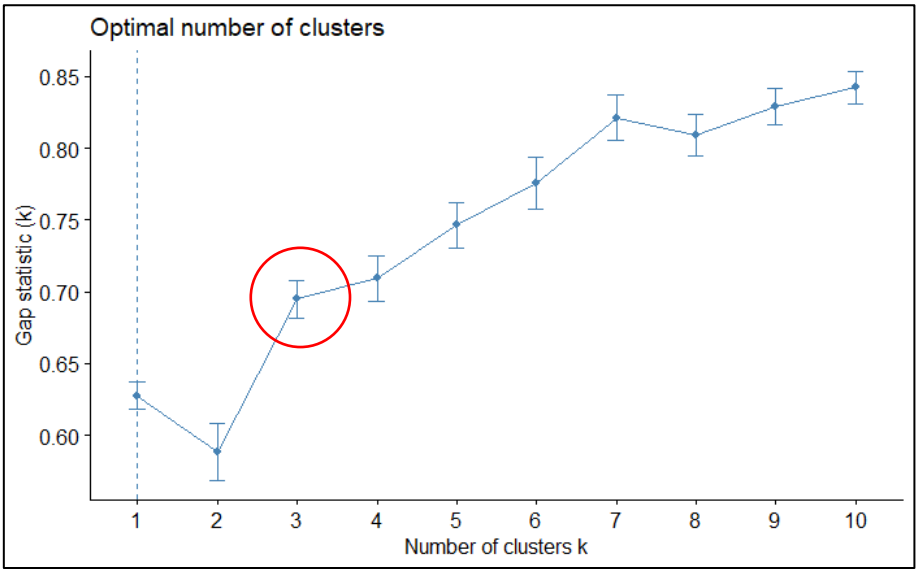

Fig. 3.2. Gap statistic plot with optimal number of clusters, k=3, highlighted with a red circle.

A sudden jump in the gap statistic plot is the indicator of the optimal number of clusters to be formed, which in our case is k=3, as shown in Fig. 3.2.

Following this, cluster analysis is conducted on the entire 'Pollution' dataset, rather than separately on datasets split down by year. This is done to keep the classification criteria consistent throughout the years within the dataset, and also to ensure that clustering is not done relatively but rather it is based on absolute characteristics of the entire dataset.

The clusters hence formed are added to the 'Pollution' dataset, against each of the records. The encoded clusters thus formed are given names based on the levels of concern, i.e. High,

Medium and Low concern, and the final dataset is then imported into Tableau, to plot the above identified hotspots.

3.4.2. Plotting the Hotspots

Once the data is imported into Tableau, a new worksheet is created. With the assistance of location data in the form of 'Latitude' and 'Longitude', the high-concern air pollution hotspots that have been identified are plotted on the map of South Asia. Alongside these high-concern areas, medium and low-concern cities are also mapped. To visually distinguish these different clusters, Tableau's 'Marks' functionalities are utilised to assign varying colours, shapes, and sizes to the markers of each segment. Furthermore, a filter by year is applied, allowing users to observe the changing pollution levels and concerns on a year-by-year basis.

3.5. Exploratory Data Analysis on the Socio-economic Data

For the next part of the analysis, the focus is shifted back to coding in Rstudio. Further analysis is done in 2 steps:

(i) Year-on-Year Pollution Progression trends
(ii) Relation with Pollution and Socio-economic Data

3.5.1. Year-on-Year Trends

'Pollution' data is used for this analysis as well. Year-on-year data is plotted to understand the trends and insights. These trend lines show variations in levels of PM10, PM2.5 and NO_2 concentration year-on-year on the overall dataset, then on country and concern level.

These trend graphs would build a great understanding of how pollution varies with population across different concern level cities and countries.

3.5.2. Relation with the socio-economic factors

Lastly, it is also vital to have an understanding of how these pollution trends relate to socio-economic factors, which is the 3rd research objective of this study (refer to section 1.3).

Various articles, such as Gogtay and Thatte (2017), Franzese and Iuliano (2018), and Senthilnathan (2019), suggest that correlation analysis is a great technique for understanding relationships between two or more variables. Additionally, while calling correlation analysis the most commonly used technique Makowski *et al.* (2020), also defined different types of correlation techniques. They also described 'Pearson's' method of correlation as the most prevalent one, which is, therefore, also used by us in this study. On the other hand, a number of publications, like Cohen *et al.* (2013), and Sarstedt and Mooi (2019) also described regression analysis as a great tool to understand relationships between dependent variables and independent variables. Here is the mathematical notation of a linear regression model:

$$y = \alpha + \beta_1 x_1 + e$$

where according to Sarstedt and Mooi (2019),

y is the dependent variable with α intercept,

x_1 is the independent variable,

β_1 is the regression coefficient, and

e is the residual term.

Accordingly, a multiple regression model would something like this:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + e$$

where x_1, x_2,…, x_n are independent variables 1,2,…,n, and β_1, β_2,…, β_n are their respective regression coefficients.

The regression analysis model processes all the records and finds a best-fit line through all the data points to present a relationship, as visualised in the below figure (Sarstedt and Mooi, 2019).

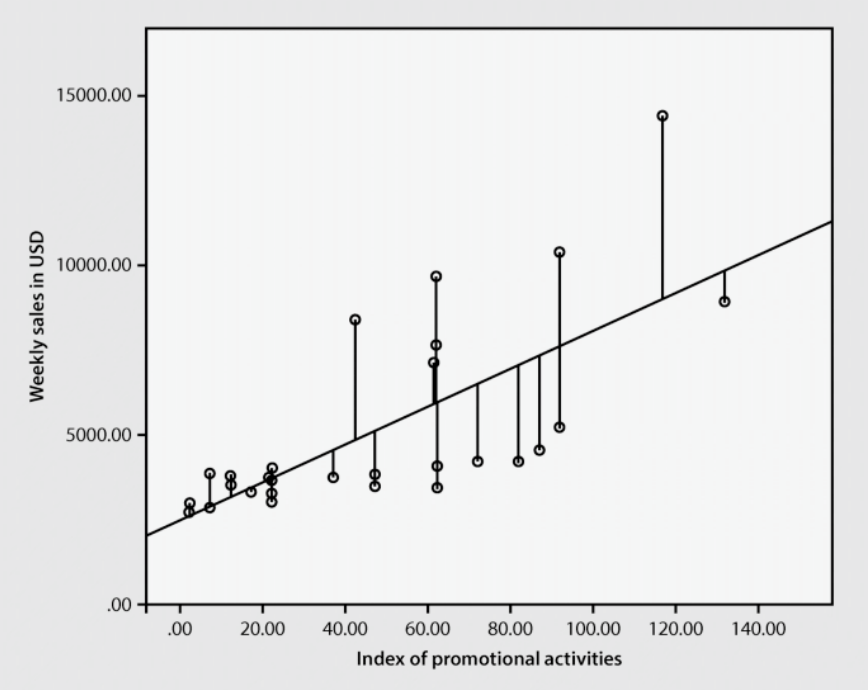

Fig. 3.3. An example to visually represent the best-fit line made by a regression model (Sarstedt and Mooi, 2019).

Therefore, for this last phase of our study, we deploy both Pearson's correlation and regression analysis to better understand the relationships in the data.

For these analyses, the 'Pollution' and socio-economic datasets are merged, since they are present in separate datasets. This is done just on the year 2020 data, since socio-economic datasets were last updated in 2020, and do not have any past data other than 2020.

Once the merged datasets are obtained, correlation matrices are generated, to understand the relationship of PM10, PM2.5 and NO_2 levels with other factors within the dataset(s). After this, regression analyses are carried out for each of the pollutants' concentration levels to better understand their relationship with the other data points.

All these analyses would help build a deep understanding of why and how the levels of pollutants are impacted by/ impacting the socio-economic factors of the region or city.

The results and interpretations are further discussed in detail in the 'Results and Discussions' chapter.

4. Chapter 4: Results and Discussion

In this chapter our focus is on interpretation and consequent discussion of the results of the analyses explained in the 'Methodology' chapter. Starting from how the clustering analysis is used in segregating the cities to identify the hotspots to understanding the relation of the pollution levels with socio-economic scenario in the South Asian cities.

4.1. Clustering analysis to find air pollution hotspots

Once the data pre-processing and cleaning are done using methods like Mahalanobis distance, we first identify the air pollution hotspots in South Asia over 10 years starting from 2010 to 2020. 'Gap Statistic' plot forms the base of clustering analysis, by identifying that ***k=3*** is the optimal number of clusters to be formed for the available data. Further, the k-means clustering method helped form clusters to differentiate cities within a continuum of concern levels. The below chunk of code is how we develop these clusters.

```
##### K-means clustering

```{r}
# Perform clustering using k-means
set.seed(10)
clusters <- kmeans(pollution_imputed_data_scaled, 3)

# Add cluster results to the data
pollution_imputed_data$cluster <- clusters$cluster
```

Fig 4.1. Using the K-means clustering method to form clusters and append them to the dataset.

The above code generates clusters that upon summarisation give below mentioned average levels of concentrations of the 3 pollutants under study.

Cluster	PM10 Levels	PM2.5 Levels	NO_2 Levels	Population	Count
1	200.24	121.07	33.44	2540164.8	221
2	64.19	38.77	16.91	638479.6	374
3	94.53	62.12	31.24	1943594.7	788

Table 4.1. Pollutants' average concentration levels across clusters hence formed.

As evident from the above table, there are very clear variations in terms of average PM10, PM2.5 and NO2 concentrations across three clusters. These clusters can be hence named as follows according to their level of concern.

1. Cluster 1 can be named ***'High Concern'***: PM10, PM2.5 and NO2 levels are highest in Cluster 1 as compared to the other clusters. These cities can also be labelled as the ***'Air Pollution hotspots***.
2. Cluster 2 can be called ***'Low Concern'***: Quite opposite to cluster 1, concentration levels in Cluster 2 are lowest across the clusters.
3. Cluster 3 can be called ***'Medium Concern'***: PM10, PM2.5 and NO2 concentration levels are higher than that of Cluster 2 but lower than that of Cluster 1.

One of the initially visible inferences, from the above summary of the 3 clusters, is that levels of pollution are directly proportional to the population of the city.

Analysing these clusters, three major observations are made:

(i) Year-wise analysis using visuals on *Tableau* (refer to Appendix IV) shows that cities farther from the coastal region are more likely to be classified as hotspots as compared to the ones closer. A few Indian and Bangladeshi cities are exceptions for being
```

classified as hotspots even after being closer to the sea. Further discussion about such cities is in the 'Conclusions' Chapter.
Note: Hotspot map is not made for year 2021, since it has only one record in the data.

(ii) Apart from the seas, another major geographical feature of South Asia is the Himalayas. The cities that lie in and around the Himalayas are mostly low or medium concern cities. For example, cities in North Pakistan, Nepal, Bhutan, and North-Eastern India are either at medium or low levels of concern. Both of these observations together indicate that the major concern lies in the Indo-Gangetic Plains, or the IGP region, which is exactly what Singh *et al.* (2021) also argued. Although the article is focused on just India, the IGP region as per one of the articles by NASA spans from Pakistan to Bangladesh (NASA, 2024).

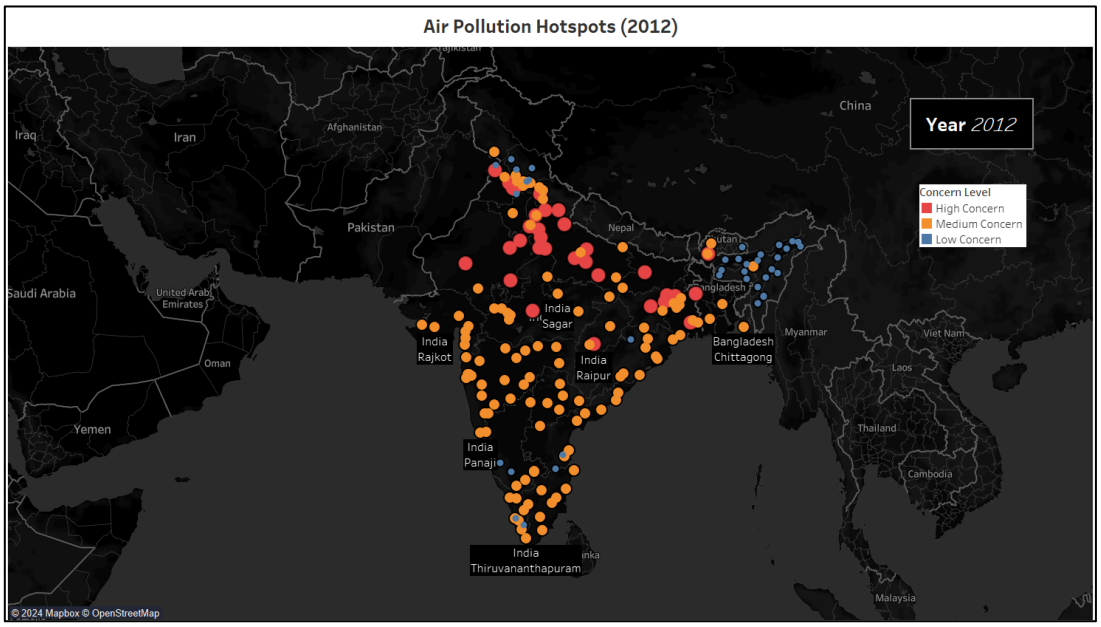

Fig 4.2. Air Pollution hotspots in the year 2012.

These inferences also justify why, for each year in our data, most of the hotspots lie in the IGP region. Fig. 4.2 visually supports the insight from our analyses.

(iii) Another interesting observation from the hotspot analysis is that even though the 2019 plot has fewer data points, 2020 has proportionately more low concern cities. Even the IGP region has relatively more low concern areas than in previous years. This sudden trend shift is most likely due to the global lockdowns during the COVID pandemic, as also discussed by Kumar *et al.* (2020).

Having understood the patterns and positions of these air pollution hotspots, we move forward to understand their trends.

4.2. Exploratory Data Analysis on Clusters

In this section, we observe some trend plots of the 'Pollution' data cut on country level, concern level and overall level. The aim is to understand the year-on-year trends. These trends are divided into 3 sub-categories, as mentioned below:

4.2.1. Year-on-year Progression on an overall level

On an overall level, year-on-year progressions of PM10 and PM2.5 are quite similar. Both of their concentration levels are stable until they spike in 2016 and 2017, and then again drop in 2018 onwards. PM10 levels also have a major drop from 2010 to 2011, which most likely is a data noise as the number of records for the year 2010 is very low for any statistical conclusion.

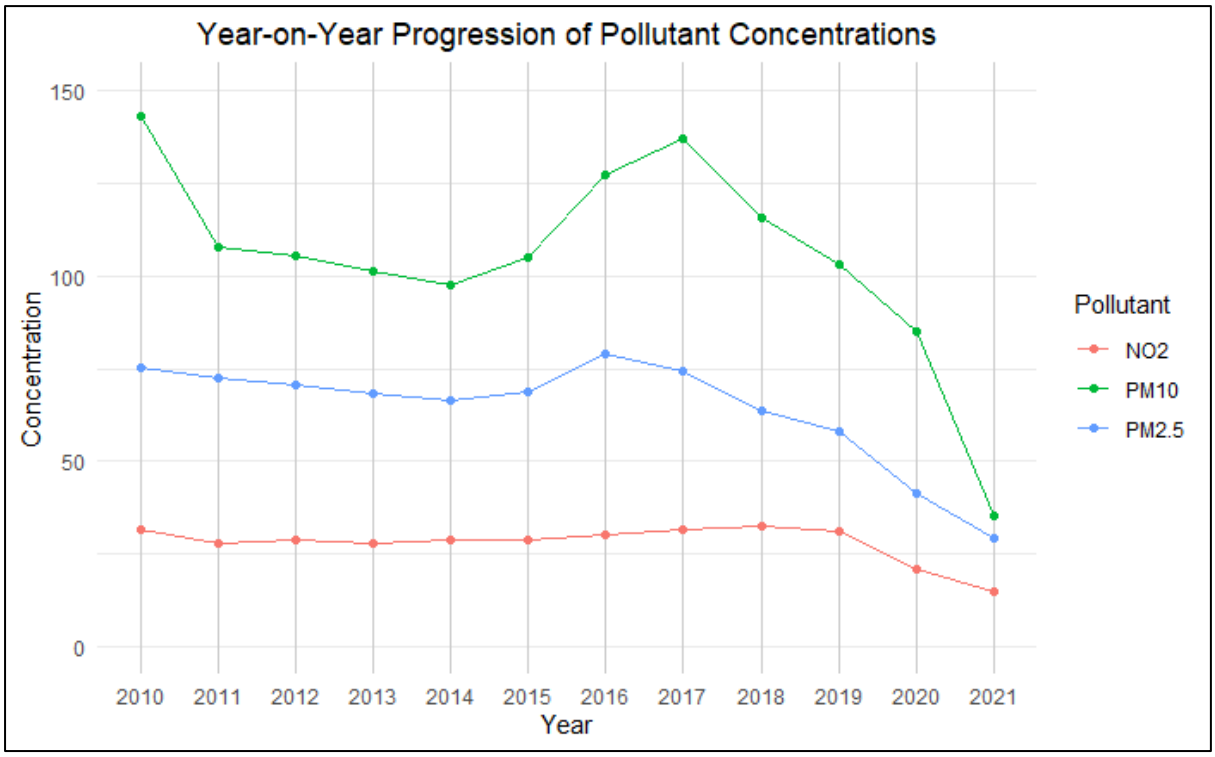

Fig 4.3. Concentration levels year-on-year.

Unlike PM10 and PM2.5 levels, average NO2 levels are consistently low and don't vary much for the entire time period. However, it does drop in 2020 potentially due to implementation of the COVID pandemic-related restrictions.

4.2.2. Year-on-year Progression at Country level

Further exploring, the country level analyses show that each country has very different trends compared to each other (for the plots refer to Appendix V). However, at the same time, they all follow the pattern of having similar trends in PM10 levels as in PM2.5 levels, especially Bangladesh, India and Pakistan. All three of them have PM10 and PM2.5 levels spiking and dipping around the same time.

Due to a larger number of available records Indian trends outweigh those of other countries, resulting in overall trends (discussed in sub-section 4.2.1.) being very similar to the Indian ones. However, it is interesting as well as slightly concerning that even though overall trends show a dip in concentration levels from 2018 onwards, PM10, PM2.5 and NO2 levels have an increasing slope from 2017 onwards in Bangladesh.

4.2.3. Year-on-year Progression on concern level

Looking at the pollution trends by levels of concern, it is observed that although there is a very small difference in average population between high and medium concern cities, PM10 and PM2.5 levels are very high in the high concern cities as compared to medium concern ones. And when low concern cities are compared, it is observed that both pollution levels and

population are relatively very low. The plots hence created for this analysis are the precise visualisations of Table 4.1 (for the plots refer to Appendix V).

After comprehending the clusters and their pollution trends, we move forward to the secondary objective of this study.

4.3. Understanding the relationship with socio-economic data

In this section, we focus on observing the relationship between the air pollution and the socio-economic environment in the South Asian cities. The relation of the pollution data is checked with the other 3 datasets and also with population data present within the pollution dataset, using the correlation matrices and linear or multiple regression analyses. Here are the results observed:

4.3.1. Relation with the 'Growth of urban areas' data

The correlation matrix, in Fig 4.4, indicates a strong positive correlation between PM10 and PM2.5 concentration levels. It also shows a very good relationship of NO_2 levels with PM10 and PM2.5 levels. However, the correlation matrix does not show any strong relationship of the PM10, PM2.5 and NO_2 concentration levels with any of the other factors within the dataset.

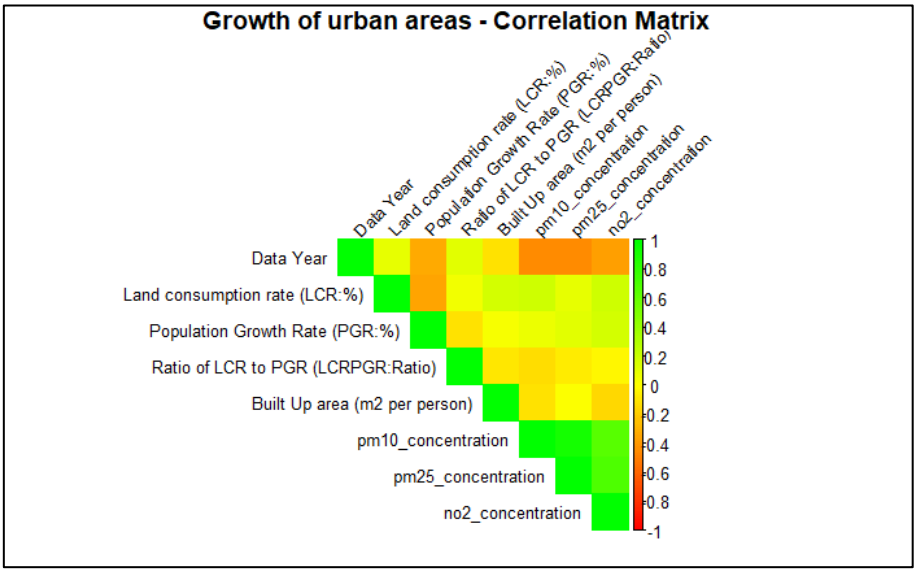

Fig 4.4. Visual representation of correlation matrix of the 'Growth of urban areas' dataset.

Regression analyses on the other hand bring up a very interesting finding.

- 'Land consumption rate' does not have any statistically significant effect, however, with every 1 unit increase it causes an average increase of 10.24 units in PM10 concentration levels ($t(25) = 0.86, p = 0.203$; CI = [-5.87, 26.34]).
- Similarly, with every 1 unit increase in it, 'Population Growth Rate' causes an average increase of 7.81 units in PM10 levels ($t(25) = 1.31, p = 0.4$; CI = [-10.98, 26.60]).

'Built Up area' on the other hand has a very minimal effect on pollutant levels. Refer to Appendix VI for regression results and other plots.

4.3.2. Relation with the 'Urban transport' data

Similar to the 'Growth of urban areas' dataset, the correlation matrix of the 'Urban transport' dataset also indicates a high correlation between PM10, PM2.5 and NO_2 concentrations with minimal correlation with any other factor. Interestingly, even the regression analyses suggest

that ‘Access to public transport’ does not have any major effect on pollution levels (refer to Appendix VI for all the plots and regression analysis). This unexpected trend can also be attributed to a couple of insights by Jain (2013):

- Many smaller South Asian towns lack organised public transport systems.
- Rapid growth of 6-10% per annum in private vehicle ownership.

These interesting findings and insights are discussed further in the ‘Conclusions’ chapter.

4.3.3. Relation with the 'Open & Green Spaces' data

Another set of insights is pulled from the 'Open & Green Spaces' dataset (refer to Appendix VI for all the plots and regression analysis):

- The correlation matrix brings up an interesting finding that the average share of open spaces for public use has a positive correlation with PM10 concentration levels.
- Even the multiple regression analysis shows a significant effect of the average share of open spaces upon PM10 concentration levels ($t(20) = 3.06, p = 0.006$), with every 1% increase in the average share of open spaces resulting in 6.54 units increase in the PM10 concentration levels (CI = [2.09, 11.00]).
- The same regression analysis also highlights a significant negative effect of the proportion of the urban population with access to open spaces on PM10 levels ($t(20) = -2.38, p = 0.028$), with every 1% of the increase in the proportion of the urban population with access to open spaces resulting in 1.67 units decrease in PM10 levels (CI = [-3.14, -0.20]).

Similarly, although not significant, the average share of open spaces for public use has a positive effect on PM2.5 and NO_2 concentration levels as well. These relationships are further discussed in the Conclusions chapter.

4.3.4. Relation with the Population data

The final set of relationship analyses is done on the ‘Pollution’ data itself. More specifically on the population data available in it. Interestingly, where most of the variables only have low correlations, latitude shows a very good correlation with PM10 concentration. This insight further strengthens the previously discussed finding that the IGP region is more likely to have a higher number of air pollution hotspots.

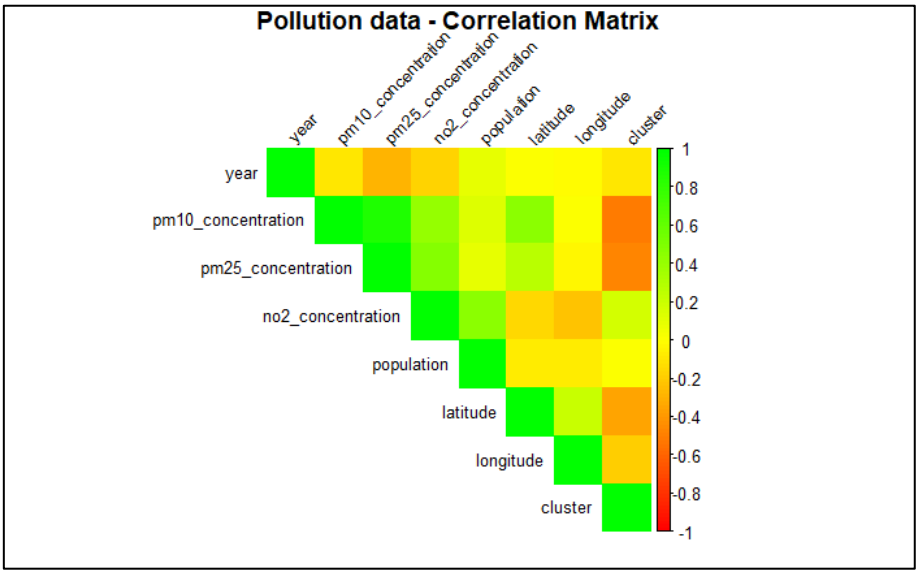

Fig 4.5. Visual representation of correlation matrix of the ‘Pollution’ dataset.

Additionally, regression analyses bring up some statistically significant relationships, as shown in the below table (refer to Appendix VI for detailed results of the regression analyses).

Coefficient	PM10 Levels	PM2.5 Levels	NO_2 Levels	Measures
Intercept	1.000e+02	6.406e+01	2.588e+01	**Estimate**
	1.582e+00	8.900e-01	2.369e-01	**Std. Error**
	63.231	71.972	109.26	**t value**
	<2e-16	<2e-16	<2e-16	**Pr(>|t|)**
population	1.903e-06	6.940e-07	1.089e-06	**Estimate**
	3.874e-07	2.180e-07	5.801e-08	**Std. Error**
	4.914	3.184	18.78	**t value**
	1e-06	0.00149	<2e-16	**Pr(>|t|)**

Table 4.2. Regression analyses results on the 'Pollution' dataset.

Here are some of the interesting observations from the above table:

- There is a significant effect of 'population' upon PM10 concentration levels ($t(1381) = 4.91, p < 0.001$). With every 1 unit increase in 'population', PM10 levels increase by 1.903e-06 units (CI = [1.143426e-06, 2.663193e-06]).
- Similarly, 'population' also has a significant effect upon PM2.5 levels ($t(1381) = 3.18, p = 0.0015$). With every 1 unit increase in 'population', PM2.5 levels increase by 6.940e-07 units (CI = [2.663389e-07, 1.121576e-06]).
- The 'population' variable also has a significant effect upon the NO_2 Levels ($t(1381) = 18.78, p < 0.001$) With every 1 unit increase in 'population', NO_2 Levels increase by 1.089e-06 units (CI = [9.754654e-07, 1.203065e-06]).

According to the above observations, although the concentration levels of pollutants rise with increasing population, the magnitude of these changes is minimal, despite being statistically significant.

All the findings and insights from this chapter are summarised as a story and further discussed in the Conclusions chapter.

5. Chapter 5: Conclusions

In this final chapter, we will summarise all the findings of this study, followed by discussions on future research that can be based on these findings.

5.1. Summarising the results

The study puts forward a very simple yet novel method of air pollution hotspot identification, i.e. the clustering approach. The approach helped identify a very interesting pattern, that the majority of hotspots in South Asia lie in the IGP or the Indo-Gangetic Plains region. This pattern is supported by a correlation analysis, which showed that latitude has a very good correlation with the PM10 concentration levels.

Since this IGP region spans through Pakistan, India and Bangladesh, as discussed in the 'Literature Review' chapter, a collaborative approach on a regional level is required, to promote and invest in clean energy, endorse eco-innovation, increase public awareness and improve regulatory framework.

The main contributors to air pollution, as discussed in the 'Literature Review' chapter, are PM10, PM2.5 and NO_2. However, it is quite evident from the trend plots and correlation analyses that PM10 and PM2.5 are more strongly associated with each other than NO_2.

Among other socio-economic factors, 'Land consumption rate' and 'Population Growth Rate' have a big impact on the PM10 concentrations. Whereas the factors like 'Access to public transport' do not have any major impact on air pollution. This can possibly be due to a lack of organised public transport services or due to rapid growth in private vehicle ownership in the region. These hypotheses can potentially be the basis for some future studies in South Asia.

5.2. Future Research

There are various other findings which could also potentially form the basis for future research in this or related areas:

- Although most of the cities in the coastal region of Southern Pakistan, Southern India, Sri Lanka, Maldives and Bangladesh are in low to medium level of concern segments, there are a few cities, particularly in India and Bangladesh which are labelled as air pollution hotspots. A number of questions may arise such as "How are these cities and their industrial/ economic activities different from other coastal and/ or IGP region cities?"
  Due to a smaller number of data points focused on such cities, these questions won't be answered by this study and would require further exploration as part of a separate future study.
- Another unusual trend identified is that the average share of open spaces for public use increases the concentration levels of pollutant types under consideration in this study, whereas ideally, it should do the exact opposite.
  - One hypothesis is that this trend could possibly be better explained if interpreted in reverse order. In simpler words, there is a possibility that these open spaces are there for public use to tackle with problem of air pollution.
  - Another hypothesis could be inadequate maintenance of these open spaces.

  These hypotheses can also form a good basis for further research on the socio-economic factors.

- Moreover, further research focused on socio-economic factors is required since most of the socio-economic datasets used in this study are 2020-based. Due to the COVID pandemic, these have limited records, however, a study with more extensive and exhaustive data could present a clearer picture and would enable a thorough exploration into detailed intricacies.
- Additionally, due to lack of consistency in different country's data repository, our study is based on a common UN data source. However, for future research, the use of a more comprehensive repository with a wider range of socio-economic datasets would be preferrable.

In conclusion, while this study has provided valuable insights into the distribution and impact of air pollution hotspots across the South Asian region, it has also highlighted several areas for further investigation. The presence of air pollution hotspots in certain coastal cities within India and Bangladesh invites a further investigation into the distinct industrial and economic factors at play. The unexpected correlation between the average share of open spaces for public use and increased pollutant levels points to the need for further research into the maintenance and intended use of these areas. Additionally, the reliance on limited socio-economic data due to the COVID pandemic calls for more comprehensive data to fully grasp the socio-economic complexities. Addressing these research gaps will be essential in developing more targeted and effective solutions to mitigate air pollution in these regions.

A. References

Aggarwal, C.C. and Reddy, C.K. (2014) 'Data clustering. Algorithms and applications. Chapman&Hall/CRC Data mining and Knowledge Discovery series, Londra.' Available at: https://charuaggarwal.net/clusterbook.pdf.

Ahmad, M., Arshad, N.I.B. and Sarlan, A.B. (2022) 'An Analysis of Students' Academic Performance Using K-Means Clustering Algorithm', in F. Saeed, F. Mohammed, and F. Ghaleb (eds) *Advances on Intelligent Informatics and Computing*. Cham: Springer International Publishing, pp. 309–318. Available at: https://doi.org/10.1007/978-3-030-98741-1_26.

Ahmed, M., Seraj, R. and Islam, S.M.S. (2020) 'The k-means Algorithm: A Comprehensive Survey and Performance Evaluation', *Electronics*, 9(8), p. 1295. Available at: https://doi.org/10.3390/electronics9081295.

Anenberg, S.C. *et al.* (2010) 'An Estimate of the Global Burden of Anthropogenic Ozone and Fine Particulate Matter on Premature Human Mortality Using Atmospheric Modeling', *Environmental Health Perspectives*, 118(9), pp. 1189–1195. Available at: https://doi.org/10.1289/ehp.0901220.

Balakrishnan, K. *et al.* (2019) 'The impact of air pollution on deaths, disease burden, and life expectancy across the states of India: the Global Burden of Disease Study 2017', *The Lancet Planetary Health*, 3(1), pp. e26–e39. Available at: https://doi.org/10.1016/S2542-5196(18)30261-4.

Chatterjee, D. *et al.* (2023) 'Source Contributions to Fine Particulate Matter and Attributable Mortality in India and the Surrounding Region', *Environmental Science & Technology*, 57(28), pp. 10263–10275. Available at: https://doi.org/10.1021/acs.est.2c07641.

Chen, H. *et al.* (2022) 'Effects of air pollution on human health – Mechanistic evidence suggested by *in vitro* and *in vivo* modelling', *Environmental Research*, 212, p. 113378. Available at: https://doi.org/10.1016/j.envres.2022.113378.

Chien, F. *et al.* (2021) 'A step toward reducing air pollution in top Asian economies: The role of green energy, eco-innovation, and environmental taxes', *Journal of Environmental Management*, 297, p. 113420. Available at: https://doi.org/10.1016/j.jenvman.2021.113420.

Cohen, J. *et al.* (2013) *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Routledge. Available at: https://doi.org/10.4324/9780203774441.

Dashdondov, K. and Kim, M.-H. (2020) 'Multivariate Outlier Removing for the Risk Prediction of Gas Leakage based Methane Gas', *Journal of the Korea Convergence Society*, 11(12), pp. 23–30. Available at: https://doi.org/10.15207/JKCS.2020.11.12.023.

Dashdondov, K. and Kim, M.-H. (2023) 'Mahalanobis Distance Based Multivariate Outlier Detection to Improve Performance of Hypertension Prediction', *Neural Processing Letters*, 55(1), pp. 265–277. Available at: https://doi.org/10.1007/s11063-021-10663-y.

Dass, A., Srivastava, S. and Chaudhary, G. (2021) 'Air pollution: A review and analysis using fuzzy techniques in Indian scenario', *Environmental Technology & Innovation*, 22, p. 101441. Available at: https://doi.org/10.1016/j.eti.2021.101441.

EEAS (2021) *South Asian Association for Regional Cooperation (SAARC) | EEAS*. Available at: https://www.eeas.europa.eu/eeas/south-asian-association-regional-cooperation-saarc_en (Accessed: 11 July 2024).

Enders, C.K. (2022) *Applied Missing Data Analysis*. Guilford Publications.

European Environment Agency (2023) *Pollution | European Environment Agency's home page*. Available at: https://www.eea.europa.eu/en/topics/in-depth/pollution?activeAccordion= (Accessed: 4 July 2024).

Flood-Garibay, J.A., Angulo-Molina, A. and Méndez-Rojas, M.Á. (2023) 'Particulate matter and ultrafine particles in urban air pollution and their effect on the nervous system', *Environmental Science: Processes & Impacts*, 25(4), pp. 704–726. Available at: https://doi.org/10.1039/D2EM00276K.

Fowler, D., Brimblecombe, P., *et al.* (2020) 'A chronology of global air quality', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 378(2183), p. 20190314. Available at: https://doi.org/10.1098/rsta.2019.0314.

Fowler, D., Pyle, J.A., *et al.* (2020) 'Global Air Quality, past present and future: an introduction', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 378(2183), p. 20190323. Available at: https://doi.org/10.1098/rsta.2019.0323.

Franzese, M. and Iuliano, A. (2018) 'Correlation analysis', in. Elsevier. Available at: https://doi.org/10.1016/B978-0-12-809633-8.20358-0.

Friedrich, M.J. (2017) 'WHO Calls Off Global Zika Emergency', *JAMA*, 317(3), p. 246. Available at: https://doi.org/10.1001/jama.2016.20447.

Gogtay, N. and Thatte, U. (2017) 'Principles of Correlation Analysis', *Journal of The Association of Physicians of India* [Preprint].

Guttikunda, S.K., Goel, R. and Pant, P. (2014) 'Nature of air pollution, emission sources, and management in the Indian cities', *Atmospheric Environment*, 95, pp. 501–510. Available at: https://doi.org/10.1016/j.atmosenv.2014.07.006.

Habibi, R. *et al.* (2017) 'An Assessment of Spatial Pattern Characterization of Air Pollution: A Case Study of CO and PM2.5 in Tehran, Iran', *ISPRS International Journal of Geo-Information*, 6(9), p. 270. Available at: https://doi.org/10.3390/ijgi6090270.

Huang, S.-Z., Sadiq, M. and Chien, F. (2021) 'The impact of natural resource rent, financial development, and urbanization on carbon emission', *Environmental Science and Pollution Research*, 30(15), pp. 42753–42765. Available at: https://doi.org/10.1007/s11356-021-16818-7.

Huang, S.-Z., Sadiq, M. and Chien, F. (2023) 'Dynamic nexus between transportation, urbanization, economic growth and environmental pollution in ASEAN countries: does environmental regulations matter?', *Environmental Science and Pollution Research*, 30(15), pp. 42813–42828. Available at: https://doi.org/10.1007/s11356-021-17533-z.

Jabbar, S.A. *et al.* (2022) 'Air Quality, Pollution and Sustainability Trends in South Asia: A Population-Based Study', *International Journal of Environmental Research and Public Health*, 19(12), p. 7534. Available at: https://doi.org/10.3390/ijerph19127534.

Jacob, D.J. and Winner, D.A. (2009) 'Effect of climate change on air quality', *Atmospheric Environment*, 43(1), pp. 51–63. Available at: https://doi.org/10.1016/j.atmosenv.2008.09.051.

Jain, A.K. (2013) 'Sustainable Urban Mobility in Southern Asia'. Available at: https://unhabitat.org/sites/default/files/2013/06/GRHS.2013.Regional.Southern.Asia_.pdf.

Jiying, W., Beraud, J.-J.D. and Xicang, Z. (2023) 'Investigating the impact of air pollution in selected African developing countries', *Environmental Science and Pollution Research*, 30(23), pp. 64460–64471. Available at: https://doi.org/10.1007/s11356-023-26998-z.

Kassambara, A. and Mundt, F. (2016) 'Factoextra: extract and visualize the results of multivariate data analyses', *R Package Version*, 1. Available at: https://rpkgs.datanovia.com/factoextra/ (Accessed: 7 August 2024).

Kieschnick, R. and McCullough, B.D. (2003) 'Regression analysis of variates observed on (0, 1): percentages, proportions and fractions', *Statistical Modelling*, 3(3), pp. 193–213. Available at: https://doi.org/10.1191/1471082X03st053oa.

Komorowski, M. *et al.* (2016) 'Exploratory Data Analysis', in MIT Critical Data (ed.) *Secondary Analysis of Electronic Health Records*. Cham: Springer International Publishing, pp. 185–203. Available at: https://doi.org/10.1007/978-3-319-43742-2_15.

Kozáková, J. *et al.* (2019) 'The influence of local emissions and regional air pollution transport on a European air pollution hot spot', *Environmental Science and Pollution Research*, 26(2), pp. 1675–1692. Available at: https://doi.org/10.1007/s11356-018-3670-y.

Kumar, D. (2012) *Genomics and Health in the Developing World*. OUP USA.

Kumar, P. *et al.* (2020) 'Temporary reduction in fine particulate matter due to "anthropogenic emissions switch-off" during COVID-19 lockdown in Indian cities', *Sustainable Cities and Society*, 62, p. 102382. Available at: https://doi.org/10.1016/j.scs.2020.102382.

Kwak, S.K. and Kim, J.H. (2017) 'Statistical data preparation: management of missing values and outliers', *Korean Journal of Anesthesiology*, 70(4), pp. 407–411. Available at: https://doi.org/10.4097/kjae.2017.70.4.407.

Lelieveld, J. *et al.* (2015) 'The contribution of outdoor air pollution sources to premature mortality on a global scale', *Nature*, 525(7569), pp. 367-371M. Available at: https://doi.org/10.1038/nature15371.

Little, R.J.A. and Rubin, D.B. (2019) *Statistical Analysis with Missing Data*. John Wiley & Sons.

Mahalanobis, P.C. (2018) 'On the Generalized Distance in Statistics', *Sankhyā: The Indian Journal of Statistics, Series A (2008-)*, 80, pp. S1–S7.

Makowski, D. *et al.* (2020) 'Methods and Algorithms for Correlation Analysis in R', *Journal of Open Source Software*, 5(51), p. 2306. Available at: https://doi.org/10.21105/joss.02306.

Manisalidis, I. *et al.* (2020) 'Environmental and Health Impacts of Air Pollution: A Review', *Frontiers in Public Health*, 8. Available at: https://doi.org/10.3389/fpubh.2020.00014.

Mehmood, K. *et al.* (2021) 'Pollution characteristics and human health risk assessments of toxic metals and particle pollutants via soil and air using geoinformation in urbanized

city of Pakistan', *Environmental Science and Pollution Research*, 28(41), pp. 58206–58220. Available at: https://doi.org/10.1007/s11356-021-14436-x.

M. Fiore, A. *et al.* (2012) 'Global air quality and climate', *Chemical Society Reviews*, 41(19), pp. 6663–6683. Available at: https://doi.org/10.1039/C2CS35095E.

Mogno, C. *et al.* (2021) 'Seasonal distribution and drivers of surface fine particulate matter and organic aerosol over the Indo-Gangetic Plain', *Atmospheric Chemistry and Physics*, 21(14), pp. 10881–10909. Available at: https://doi.org/10.5194/acp-21-10881-2021.

Mohsin, M. *et al.* (2021) 'Assessing the impact of transition from nonrenewable to renewable energy consumption on economic growth-environmental nexus from developing Asian economies', *Journal of Environmental Management*, 284, p. 111999. Available at: https://doi.org/10.1016/j.jenvman.2021.111999.

Mor, S. and Ghimire, M. (2022) 'Transparency and Nationally Determined Contributions: A Review of the Paris Agreement', 11, pp. 106–119.

Moran, D. and Kanemoto, K. (2016) 'Tracing global supply chains to air pollution hotspots', *Environmental Research Letters*, 11(9), p. 094017. Available at: https://doi.org/10.1088/1748-9326/11/9/094017.

Murtagh, F. and Contreras, P. (2017) 'Algorithms for hierarchical clustering: an overview, II', *WIREs Data Mining and Knowledge Discovery*, 7(6), p. e1219. Available at: https://doi.org/10.1002/widm.1219.

NASA (2024) *Fog Blankets the Indo-Gangetic Plain*. NASA Earth Observatory. Available at: https://earthobservatory.nasa.gov/images/152337/fog-blankets-the-indo-gangetic-plain (Accessed: 12 August 2024).

National Geographic (2024) *Pollution*. Available at: https://education.nationalgeographic.org/resource/pollution (Accessed: 4 July 2024).

NIEHS (2023) *Air Pollution and Your Health*, *National Institute of Environmental Health Sciences*. Available at: https://www.niehs.nih.gov/health/topics/agents/air-pollution (Accessed: 4 July 2024).

Páez, A. and Boisjoly, G. (2022) 'Exploratory Data Analysis', in Páez, A. and Boisjoly, G., *Discrete Choice Analysis with R*. Cham: Springer International Publishing (Use R!), pp. 25–64. Available at: https://doi.org/10.1007/978-3-031-20719-8_2.

Ritchie, H. and Roser, M. (2024) 'Air Pollution', *Our World in Data* [Preprint]. Available at: https://ourworldindata.org/air-pollution (Accessed: 4 July 2024).

Rodrigues, J.N., Bhattacharya, S. and Cabete, D.C.R. (2023) 'THE IMPACT OF URBANISATION ON LONG-TERM SUSTAINABILITY IN SOUTH ASIA', *Novos Estudos Jurí-dicos*, 28(3), pp. 642–667. Available at: https://doi.org/10.14210/nej.v28n3.p642-667.

Rokach, L. and Maimon, O. (2005) 'Clustering Methods', in O. Maimon and L. Rokach (eds) *Data Mining and Knowledge Discovery Handbook*. Boston, MA: Springer US, pp. 321–352. Available at: https://doi.org/10.1007/0-387-25465-X_15.

Röser, F. *et al.* (2020) 'Ambition in the making: analysing the preparation and implementation process of the Nationally Determined Contributions under the Paris Agreement',

Climate Policy, 20(4), pp. 415–429. Available at: https://doi.org/10.1080/14693062.2019.1708697.

Rousseeuw, P.J. and Hubert, M. (2011) 'Robust statistics for outlier detection', *WIREs Data Mining and Knowledge Discovery*, 1(1), pp. 73–79. Available at: https://doi.org/10.1002/widm.2.

Roy, S. *et al.* (2023) 'Impact of fine particulate matter and toxic gases on the health of school children in Dhaka, Bangladesh', *Environmental Research Communications*, 5(2), p. 025004. Available at: https://doi.org/10.1088/2515-7620/acb90d.

Rubin, D.B. (1996) 'Multiple Imputation after 18+ Years', *Journal of the American Statistical Association*, 91(434), pp. 473–489. Available at: https://doi.org/10.1080/01621459.1996.10476908.

Salgado, C.M. *et al.* (2016) 'Missing Data', in Mit Critical Data, *Secondary Analysis of Electronic Health Records*. Cham: Springer International Publishing, pp. 143–162. Available at: https://doi.org/10.1007/978-3-319-43742-2_13.

Sarstedt, M. and Mooi, E. (2019) 'Regression Analysis', in M. Sarstedt and E. Mooi (eds) *A Concise Guide to Market Research: The Process, Data, and Methods Using IBM SPSS Statistics*. Berlin, Heidelberg: Springer, pp. 209–256. Available at: https://doi.org/10.1007/978-3-662-56707-4_7.

Savaresi, A. (2016) 'The Paris Agreement: a new beginning?', *Journal of Energy & Natural Resources Law*, 34(1), pp. 16–26. Available at: https://doi.org/10.1080/02646811.2016.1133983.

Senthilnathan, S. (2019) 'Usefulness of Correlation Analysis', *SSRN Electronic Journal* [Preprint]. Available at: https://doi.org/10.2139/ssrn.3416918.

Shi, P. *et al.* (2019) 'Data Consistency Theory and Case Study for Scientific Big Data', *Information*, 10(4), p. 137. Available at: https://doi.org/10.3390/info10040137.

Silva, R.A. *et al.* (2013) 'Global premature mortality due to anthropogenic outdoor air pollution and the contribution of past climate change', *Environmental Research Letters*, 8(3), p. 034005. Available at: https://doi.org/10.1088/1748-9326/8/3/034005.

Singh, N. *et al.* (2021) 'Air Pollution Over India: Causal Factors for the High Pollution with Implications for Mitigation', *ACS Earth and Space Chemistry*, 5(12), pp. 3297–3312. Available at: https://doi.org/10.1021/acsearthspacechem.1c00170.

Stekhoven, D.J. and Bühlmann, P. (2012) 'MissForest—non-parametric missing value imputation for mixed-type data', *Bioinformatics*, 28(1), pp. 112–118. Available at: https://doi.org/10.1093/bioinformatics/btr597.

Theron, L.C. *et al.* (2021) 'Effects of pollution on adolescent mental health: a systematic review protocol', *Systematic Reviews*, 10(1), p. 85. Available at: https://doi.org/10.1186/s13643-021-01639-z.

Tiwari, K. *et al.* (2007) 'Selecting the Appropriate Outlier Treatment for Common Industry Applications'.

United Nations (2015) *The Paris Agreement*, *United Nations*. United Nations. Available at: https://www.un.org/en/climatechange/paris-agreement (Accessed: 6 July 2024).

United Nations Climate Change (2021) *What is the Kyoto Protocol? | UNFCCC*. Available at: https://unfccc.int/kyoto_protocol (Accessed: 7 July 2024).

US EPA, O. (2016) *Particulate Matter (PM) Basics*. Available at: https://www.epa.gov/pm-pollution/particulate-matter-pm-basics (Accessed: 4 July 2024).

Wang, Q. (2018) 'Urbanization and Global Health: The Role of Air Pollution', *Iranian Journal of Public Health*, 47(11), pp. 1644–1652.

Weaver, C.P. *et al.* (2009) 'A Preliminary Synthesis of Modeled Climate Change Impacts on U.s. Regional Ozone Concentrations', *Bulletin of the American Meteorological Society*, 90(12), pp. 1843–1863.

West, J.J. *et al.* (2013) 'Co-benefits of mitigating global greenhouse gas emissions for future air quality and human health', *Nature Climate Change*, 3(10), pp. 885–889. Available at: https://doi.org/10.1038/nclimate2009.

WHO (2014) *7 million premature deaths annually linked to air pollution*. Available at: https://www.who.int/news/item/25-03-2014-7-million-premature-deaths-annually-linked-to-air-pollution (Accessed: 9 July 2024).

Wierzchoń, S. and Kłopotek, M. (2018) *Modern Algorithms of Cluster Analysis*. Cham: Springer International Publishing (Studies in Big Data). Available at: https://doi.org/10.1007/978-3-319-69308-8.

World Health Organization (2019) *Air pollution*. Available at: https://www.who.int/health-topics/air-pollution#tab=tab_2 (Accessed: 4 July 2024).

World Health Organization (2022) *Ambient (outdoor) air pollution*. Available at: https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health (Accessed: 8 July 2024).

World Health Organization (2024) *Air quality database*. Available at: https://www.who.int/data/gho/data/themes/air-pollution/who-air-quality-database (Accessed: 8 July 2024).

WorldBank (2024) *Overview*, *World Bank*. Available at: https://www.worldbank.org/en/region/sar/overview (Accessed: 11 July 2024).

Xu, R. and Wunsch, D. (2005) 'Survey of clustering algorithms', *IEEE Transactions on Neural Networks*, 16(3), pp. 645–678. Available at: https://doi.org/10.1109/TNN.2005.845141.

Zhan, C. *et al.* (2023) 'Impacts of urbanization on air quality and the related health risks in a city with complex terrain', *Atmospheric Chemistry and Physics*, 23(1), pp. 771–788. Available at: https://doi.org/10.5194/acp-23-771-2023.

Zhang, J. *et al.* (2024) 'Adverse effects of exposure to fine particles and ultrafine particles in the environment on different organs of organisms', *Journal of Environmental Sciences*, 135, pp. 449–473. Available at: https://doi.org/10.1016/j.jes.2022.08.013.

Zhang, Y., Hannigan, M. and Lv, Q. (2021) 'Air Pollution Hotspot Detection and Source Feature Analysis using Cross-Domain Urban Data', in *Proceedings of the 29th International Conference on Advances in Geographic Information Systems*. *SIGSPATIAL '21: 29th International Conference on Advances in Geographic Information Systems*, Beijing China: ACM, pp. 592–595. Available at: https://doi.org/10.1145/3474717.3484263.

B. Appendix

Appendix I.

Data sources:

(i) Pollution Data: 'Air quality database 2022 (V5)' dataset; *Accessed: 6th June, 2024.* (https://www.who.int/publications/m/item/who-ambient-air-quality-database-(update-jan-2024))

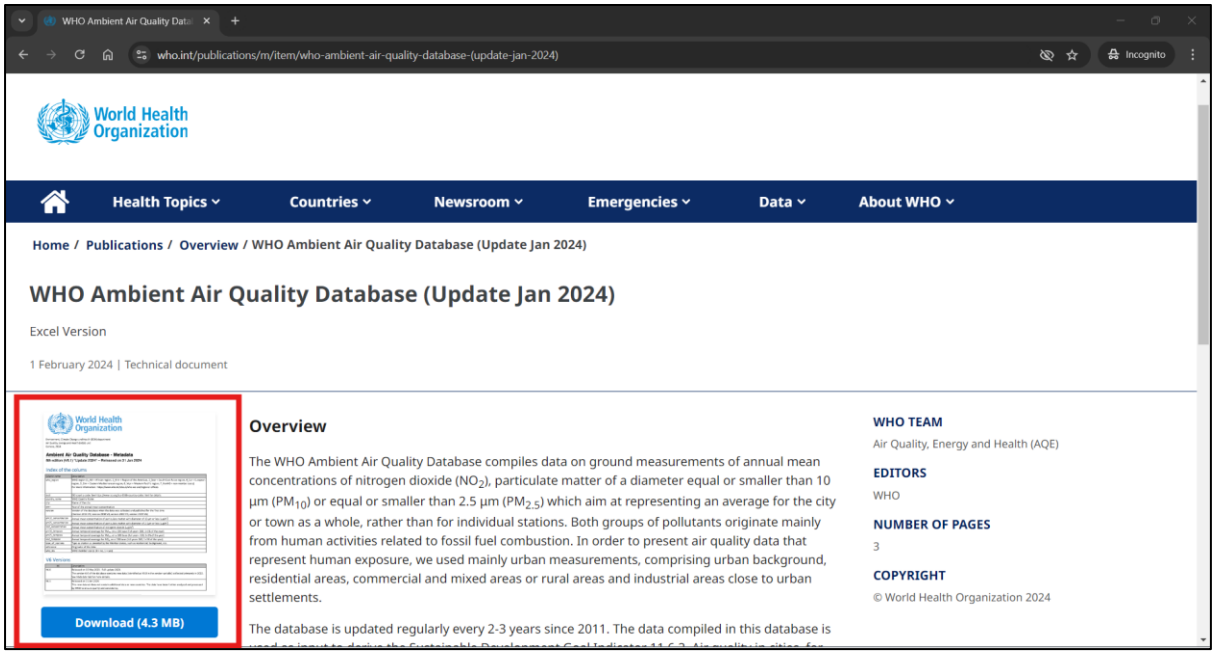

Fig B.1. Red highlighter box locates the dataset on the above mentioned webpage.

(ii) Socio-economic Data

a. 'Open & Green Spaces' dataset; *Accessed: 13th July, 2024.* (https://data.unhabitat.org/pages/open-spaces-and-green-areas)

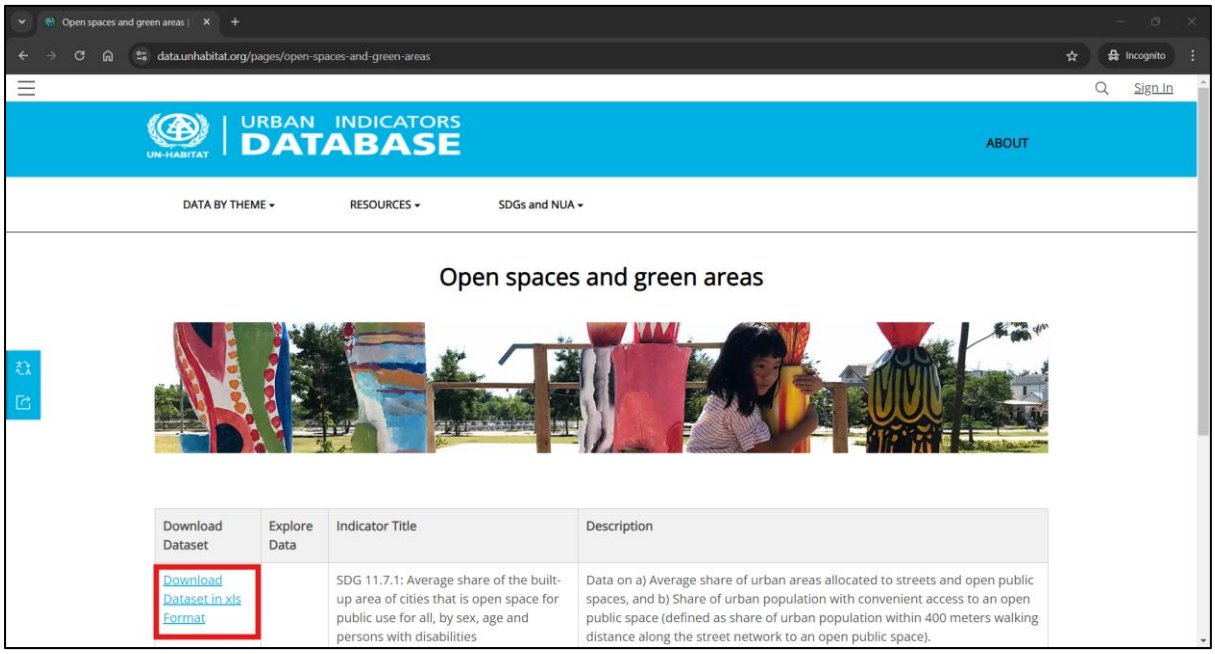

Fig B.2. Red highlighter box locates the dataset on the above mentioned webpage.

b. ‘Urban transport’ dataset; *Accessed: 13th July, 2024.* (https://data.unhabitat.org/pages/urban-transport)

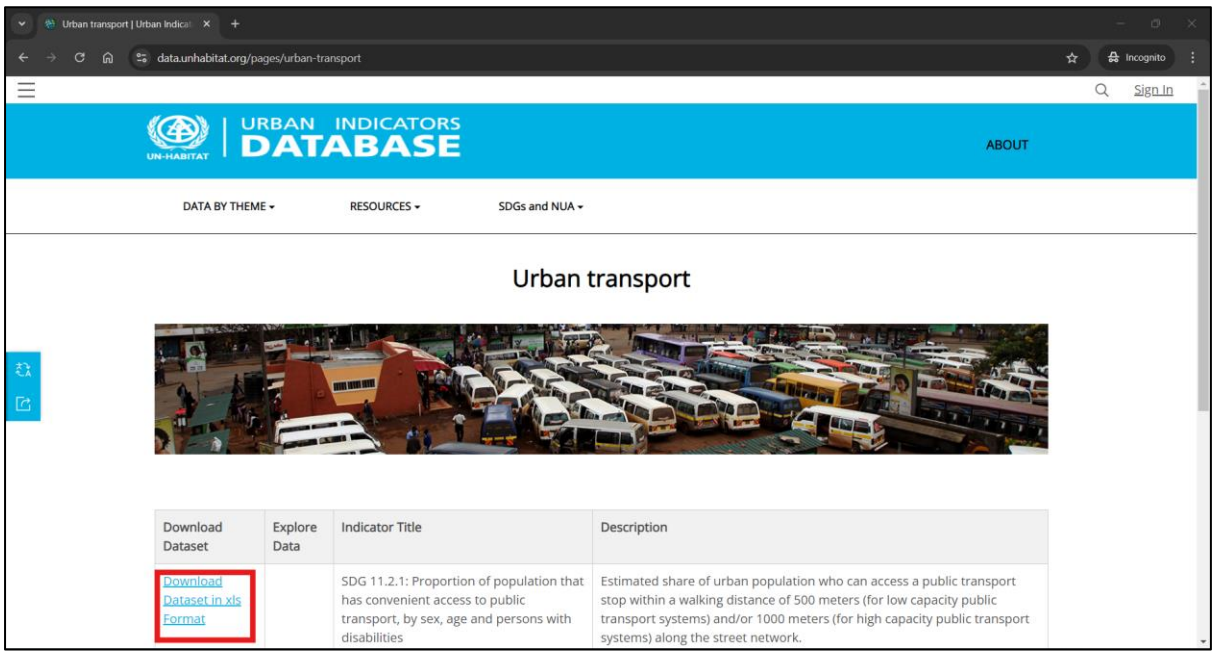

Fig B.3. Red highlighter box locates the dataset on the above mentioned webpage.

c. ‘Growth of urban areas’ dataset; *Accessed: 13th July, 2024.* (https://data.unhabitat.org/pages/spatial-growth-of-cities-and-urban-areas)

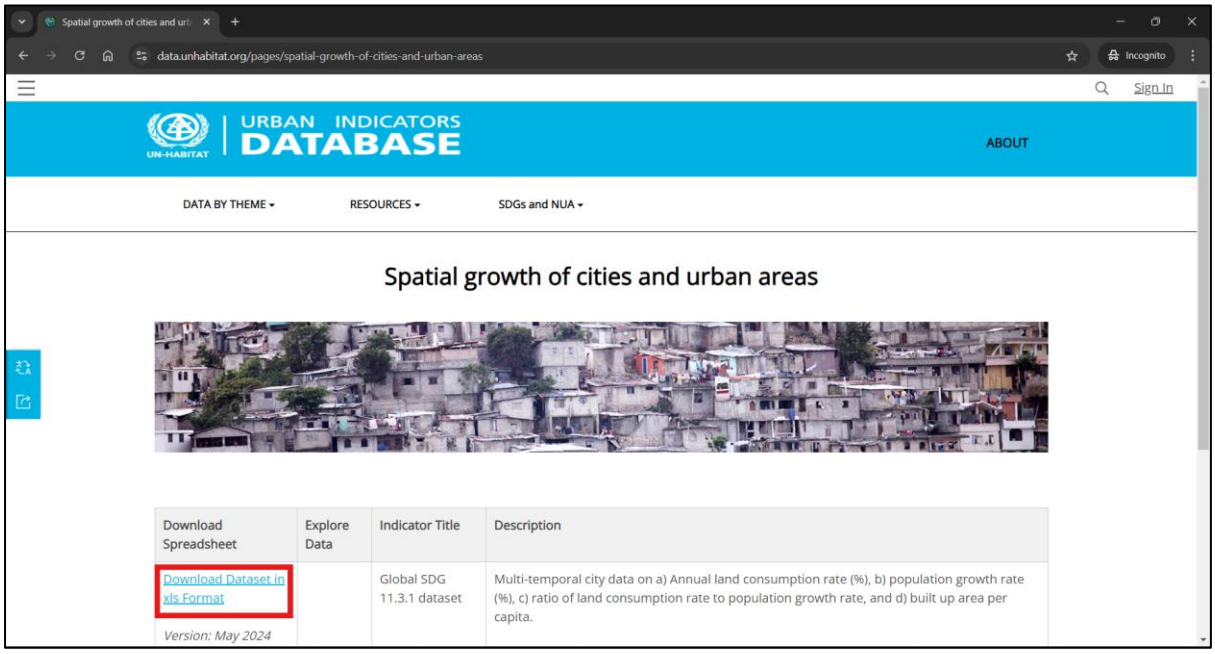

Fig B.4. Red highlighter box locates the dataset on the above mentioned webpage.

Appendix II.

Description and Rationale for removal of variables that are removed from the raw data files.

(i) 'Pollution' dataset

Variable Name	Description	Rationale for Removal
who_region	WHO region	Irrelevant since South Asia is the only region under study
iso3	ISO country code	'country_name' column better identifies the country
version	Version of the database when the data was first collected	Does not give any useful information for analyses
pm10_tempcov	Annual temporal coverage for PM_{10}	Does not give any useful information for analyses
pm25_tempcov	Annual temporal coverage for $PM_{2.5}$	Does not give any useful information for analyses
no2_tempcov	Annual temporal coverage for NO_2	Does not give any useful information for analyses
type_of_stations	Type as station as provided by the Member states, such as residential, background, etc.	Station is anyways representative of the entire city so not relevant
reference	Originator of the data	Does not give any useful information for analyses
web_link	Link to web source of the data	Does not give any useful information for analyses
population_source	Source of the population estimate	Does not give any useful information for analyses
who_ms	WHO member states	Irrelevant since all the SAARC countries are anyways analysed

Table B.1. Rational for removal of removed variables in the 'Pollution' dataset.

(ii) 'Open & Green Spaces' dataset

Variable Name	Rationale for Removal
SDG Goal	Same value for all records
SDG Target	Same value for all records
SDG Indicator	Same value for all records
Country or Territory Code	'Country or Territory Name' column better identifies the country
SDG Region	Irrelevant since South Asia is the only region under study
SDG Sub-Region	Irrelevant since South Asia is the only region under study
City Code	'City Name' column better identifies the country
Data Units	Same value for all records
Data Source	Does not give any useful information for analyses
FootNote	Does not give any useful information for analyses

Table B.2. Rational for removal of removed variables in the 'Open & Green Spaces' dataset.

(iii) ‘Urban transport’ dataset

Variable Name	Rationale for Removal
SDG Goal	Same value for all records
SDG Target	Same value for all records
SDG Indicator	Same value for all records
Country or Territory Code	‘Country or Territory Name’ column better identifies the country
SDG Region	Irrelevant since South Asia is the only region under study
SDG Sub-Region	Irrelevant since South Asia is the only region under study
City Code	‘City Name’ column better identifies the country
Data Units	Same value for all records
Data Source	Does not give any useful information for analyses
FootNote	Does not give any useful information for analyses

Table B.3. Rational for removal of removed variables in the ‘Urban transport’ dataset.

(iv) ‘Growth of urban areas’ dataset

Variable Name	Rationale for Removal
SDG Goal	Same value for all records
SDG Target	Same value for all records
SDG Indicator	Same value for all records
Country or Territory Code	‘Country or Territory Name’ column better identifies the country
SDG Region	Irrelevant since South Asia is the only region under study
SDG Sub-Region	Irrelevant since South Asia is the only region under study
City Code	‘City Name’ column better identifies the country
Data Source	Does not give any useful information for analyses
FootNote	Does not give any useful information for analyses

Table B.4. Rational for removal of removed variables in the ‘Growth of urban areas’ dataset.

Appendix III.

‘Built Up area’ data for the potentially identified outliers from the 'Growth of urban areas' dataset.

Country Name	City Name	Built Up area Per Capita Year 1	Built Up area Per Capita Year 2	Built Up area Per Capita Year 3
Afghanistan	Kandahar	222.82	155.83	107.32
Afghanistan	Lashkargah	200.29	174.12	143.36
Bhutan	Thimphu	143.09	115.08	149.42
Nepal	Butwal	172.55	212.82	218.10
Nepal	Dharan	482.79	526.07	567.23
Sri Lanka	Embilipitiya	381.45	358.29	323.20
Sri Lanka	Hambantota	181.94	270.77	273.63
Sri Lanka	Badulla	243.04	258.81	247.97
Sri Lanka	Chilaw	187.02	190.74	197.47
Nepal	Bharatpur	130.75	162.12	172.62
Sri Lanka	Matara	120.22	124.30	181.84
Sri Lanka	Jaffna	88.70	140.51	218.38
Sri Lanka	Anuradhapura	243.04	258.81	247.97

Table B.5. Consistently high Built up area across years in the ‘Growth of urban areas’ dataset.

Appendix IV.

Year-wise plots of levels of concern.

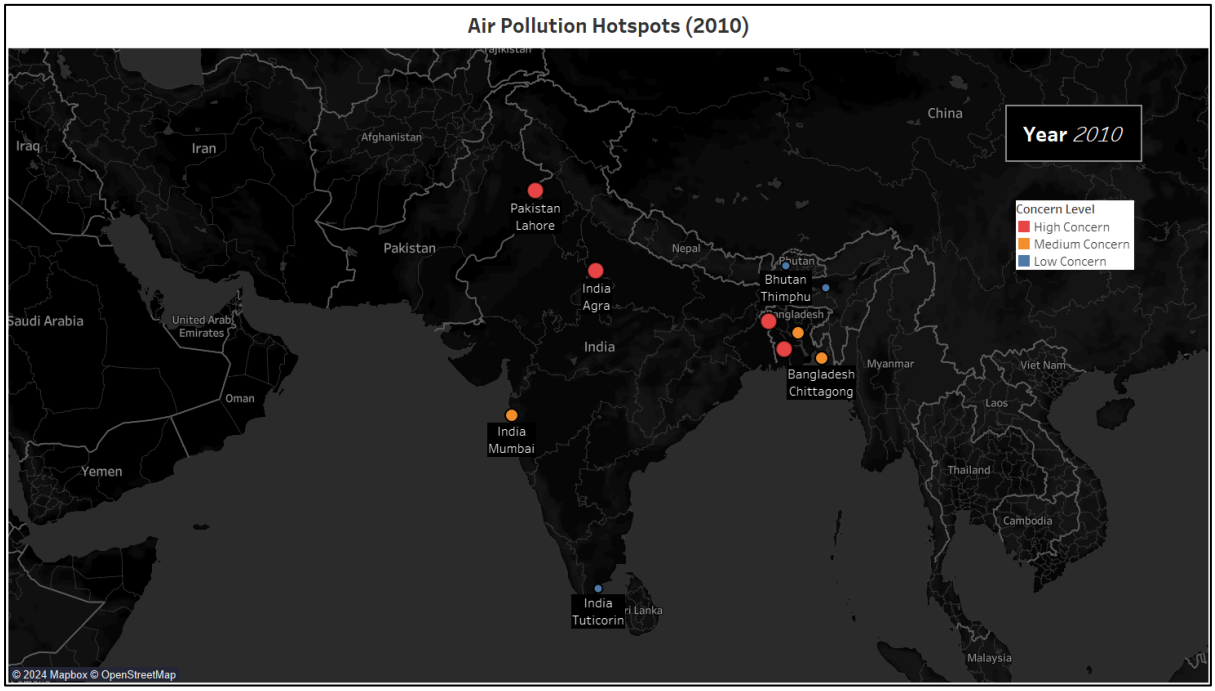

Fig B.5. 2010 hotspots plotted on the map of South Asia.

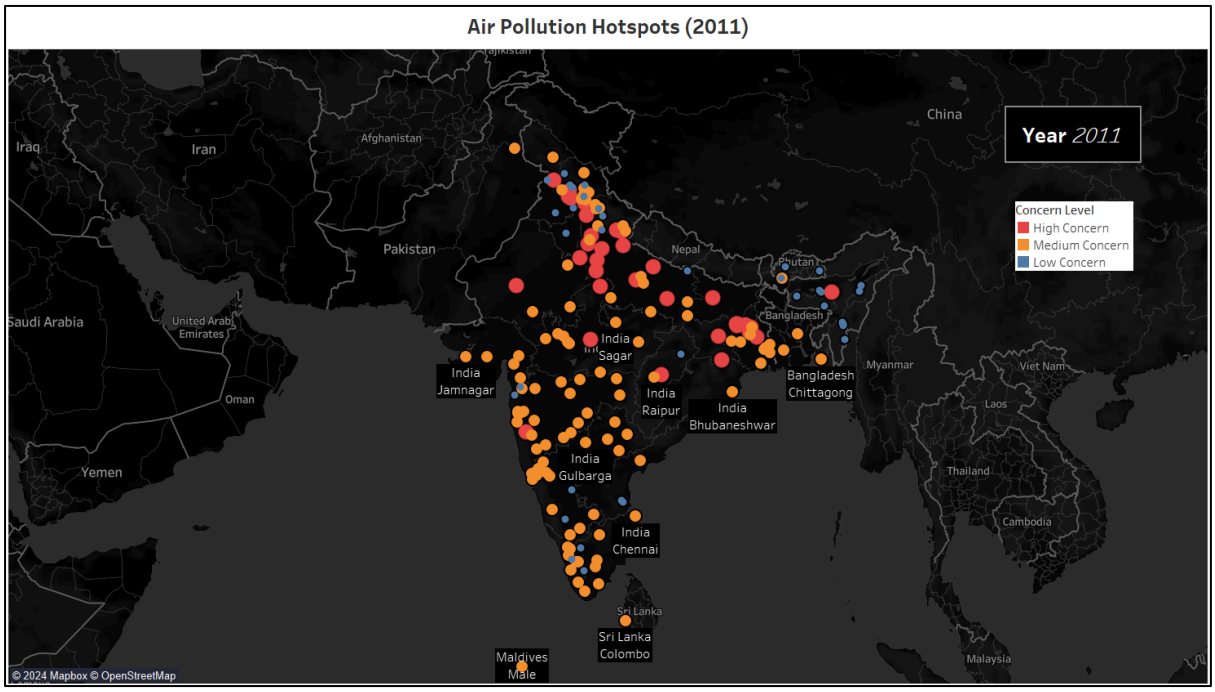

Fig B.6. 2011 hotspots plotted on the map of South Asia.

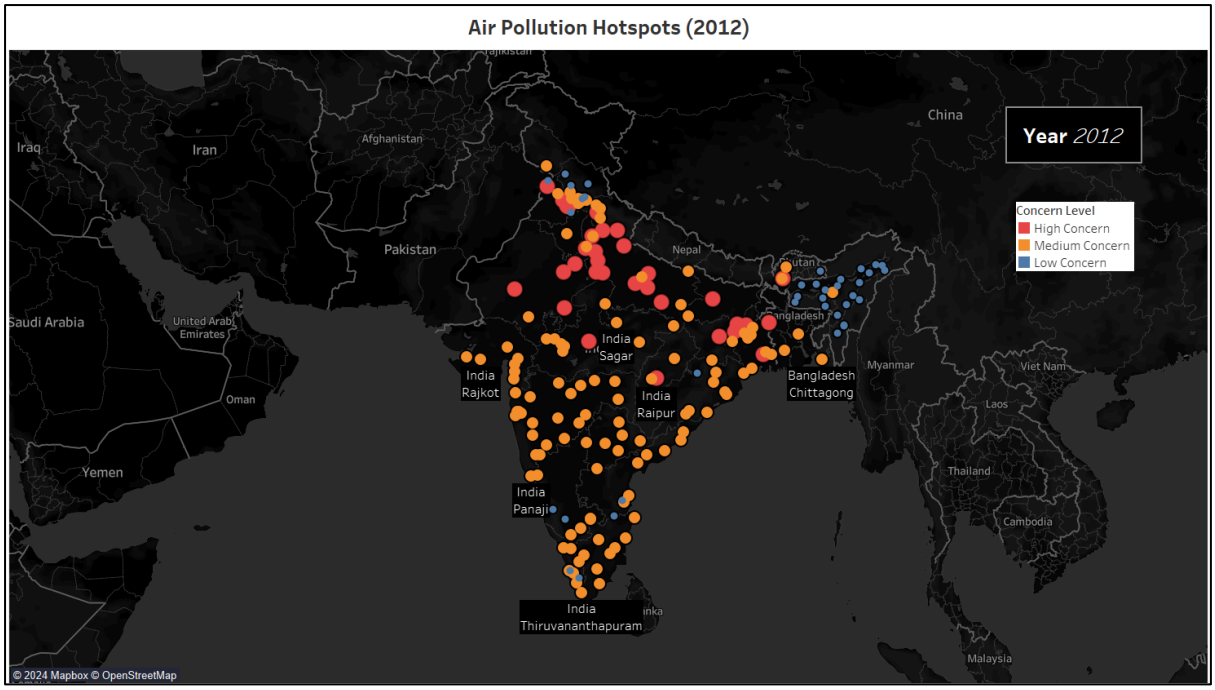

Fig B.7. 2012 hotspots plotted on the map of South Asia.

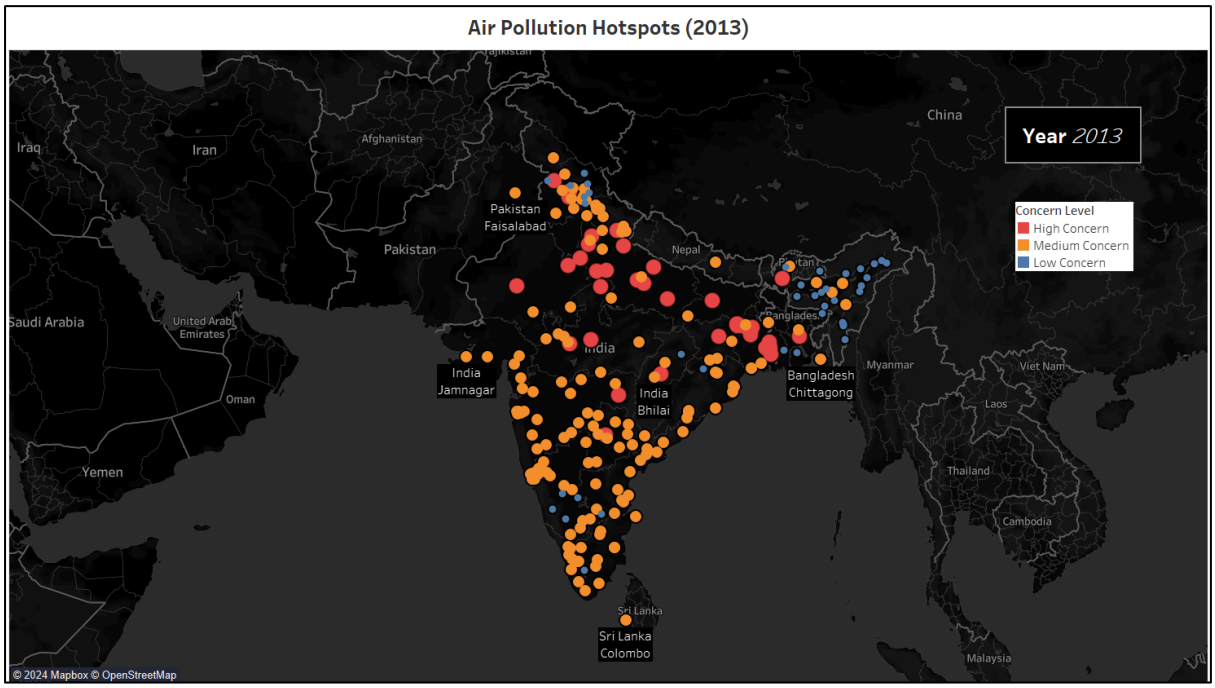

Fig B.8. 2013 hotspots plotted on the map of South Asia.

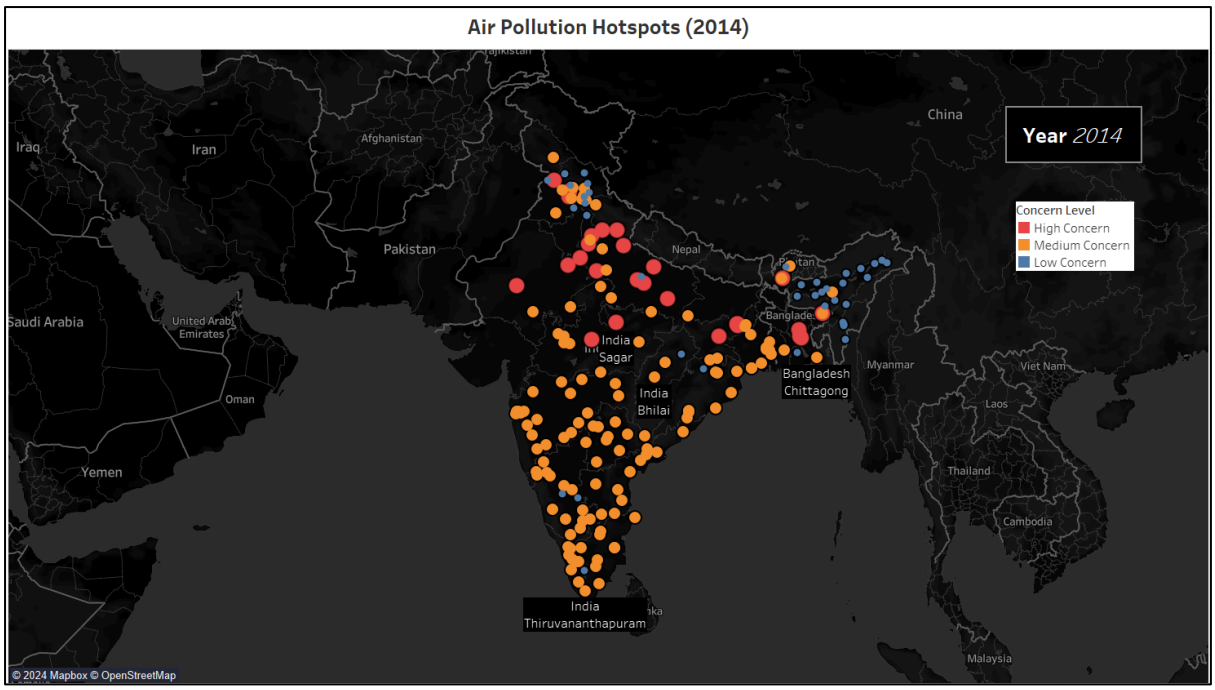

Fig B.9. 2014 hotspots plotted on the map of South Asia.

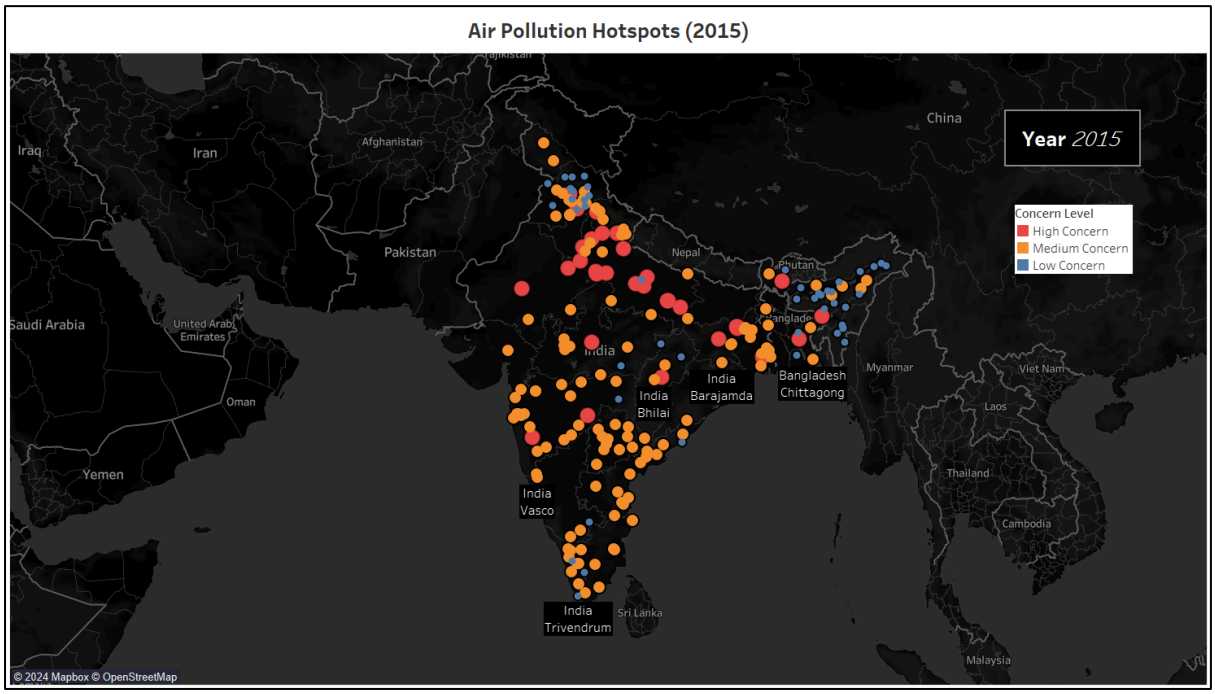

Fig B.10. 2015 hotspots plotted on the map of South Asia.

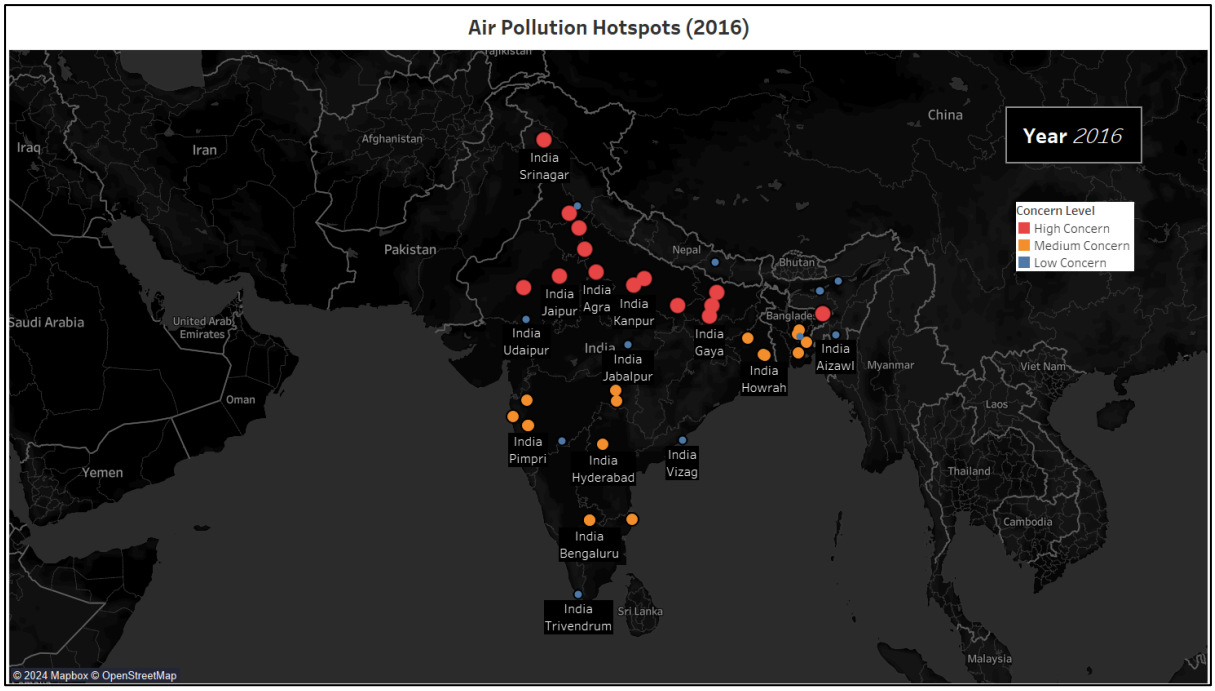

Fig B.11. 2016 hotspots plotted on the map of South Asia.

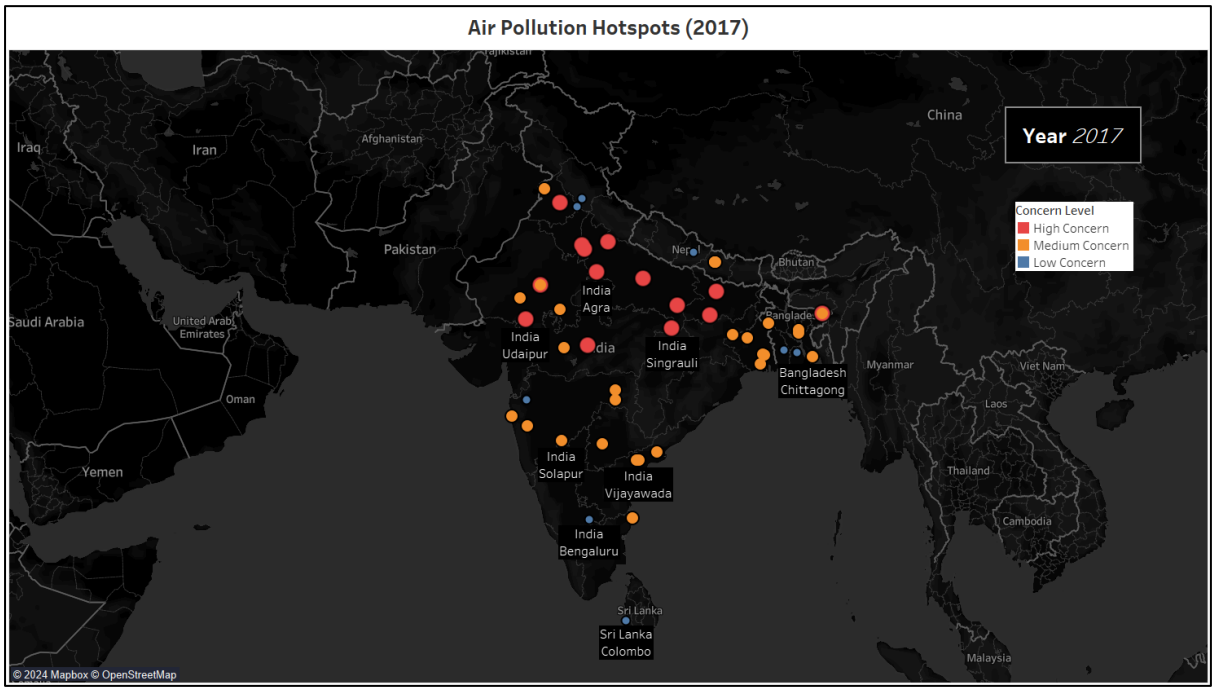

Fig B.12. 2017 hotspots plotted on the map of South Asia.

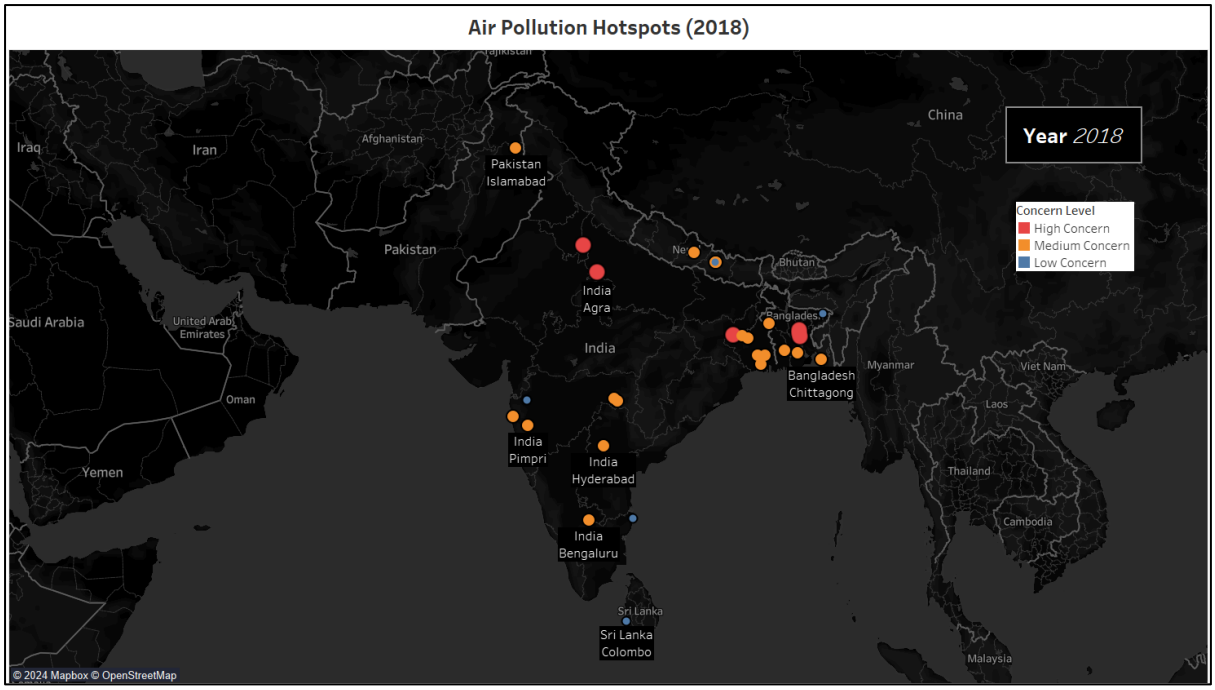

Fig B.13. 2018 hotspots plotted on the map of South Asia.

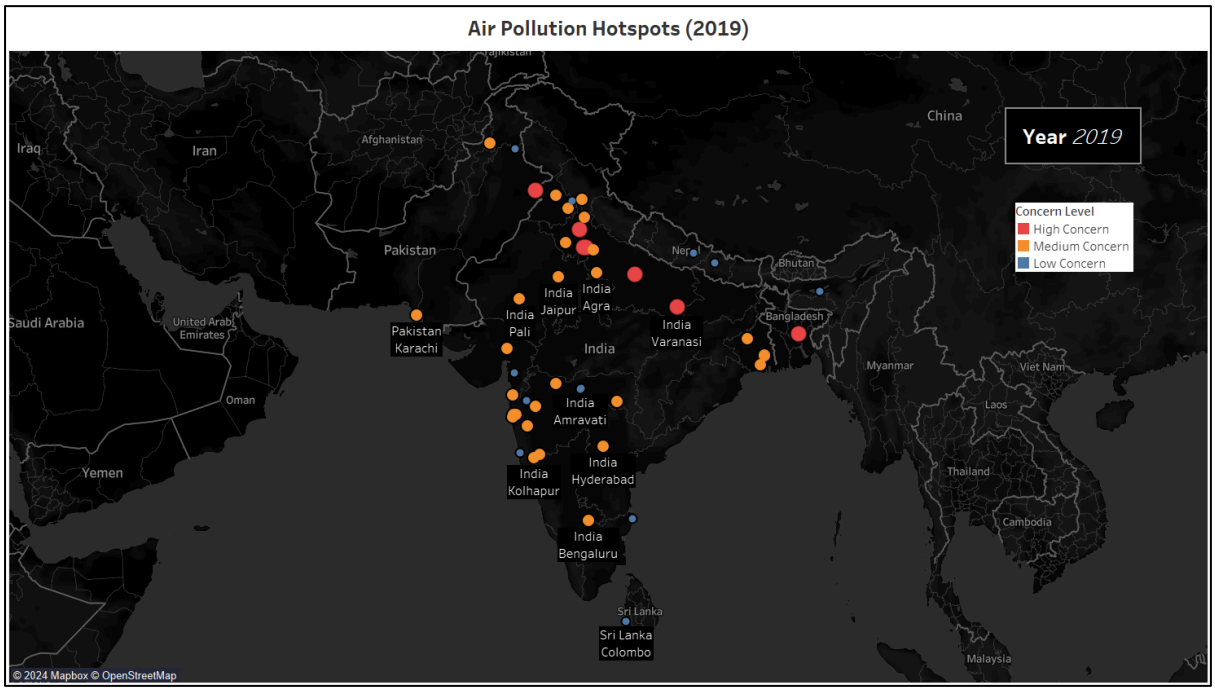

Fig B.14. 2019 hotspots plotted on the map of South Asia.

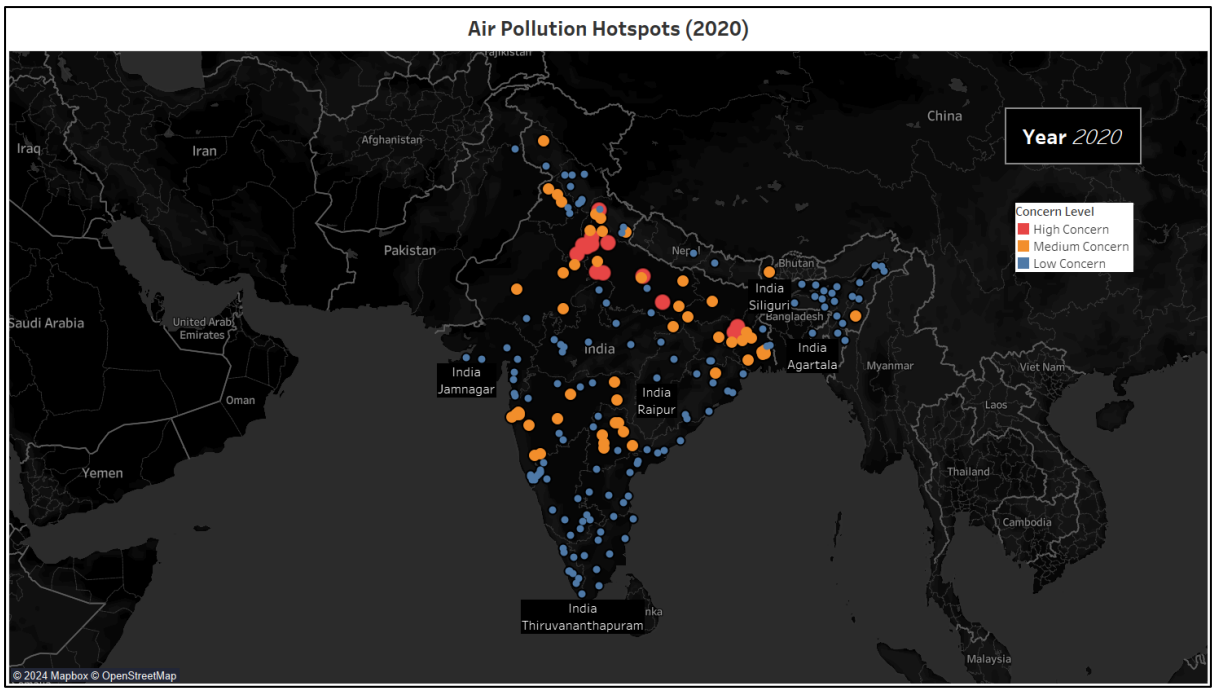

Fig B.15. 2020 hotspots plotted on the map of South Asia.

Appendix V.

Plots for section ‘4.2.2. Year-on-year Progression at country level’.

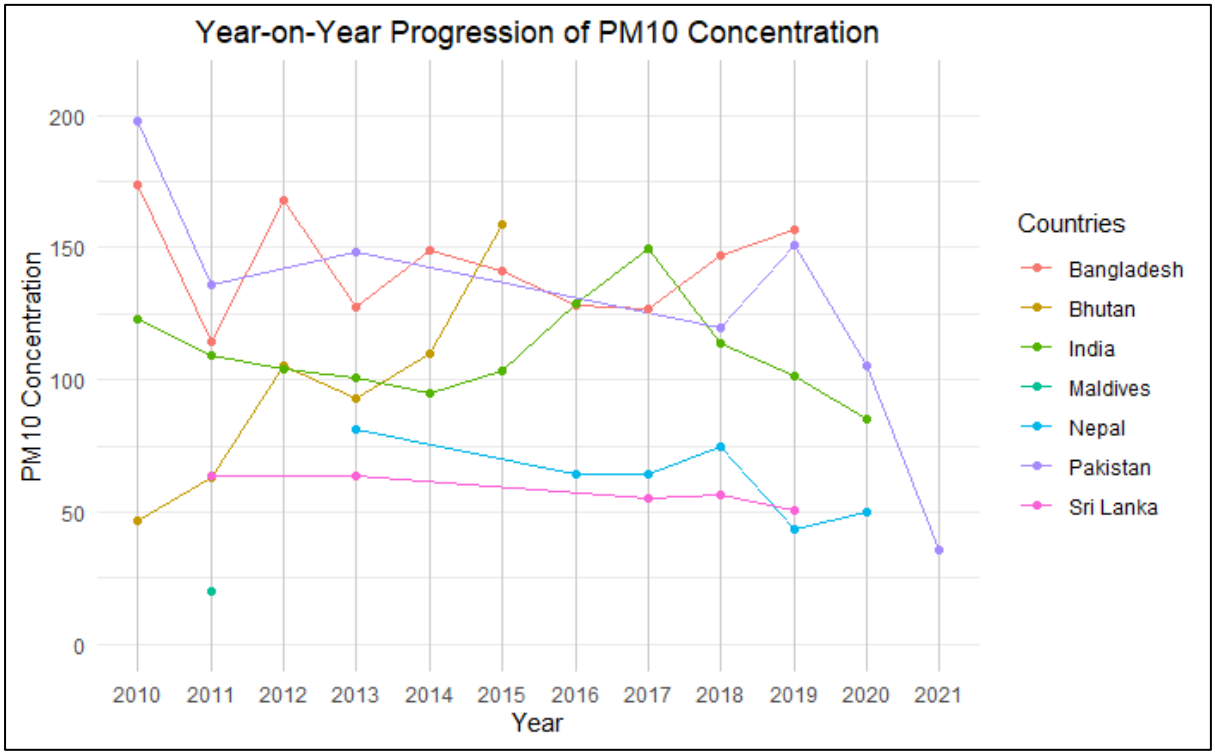

Fig B.16. Year-on-year PM10 Progression by country.

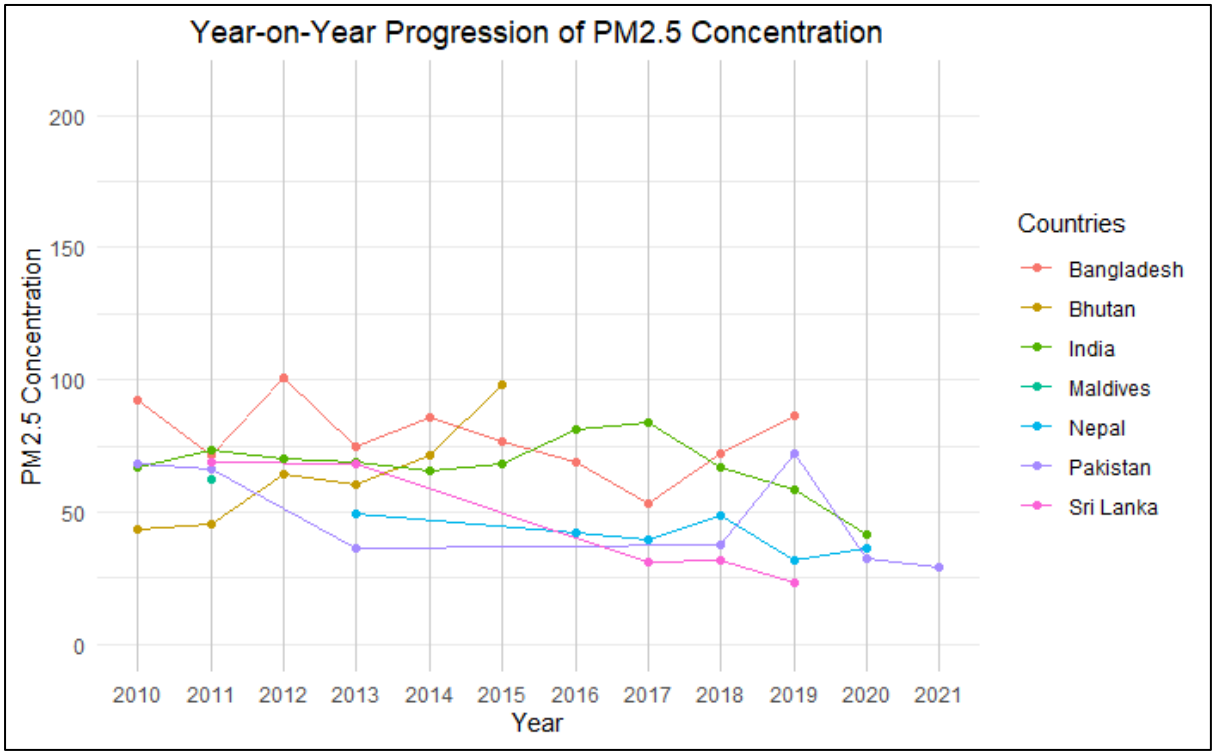

Fig B.17. Year-on-year PM2.5 Progression by country.

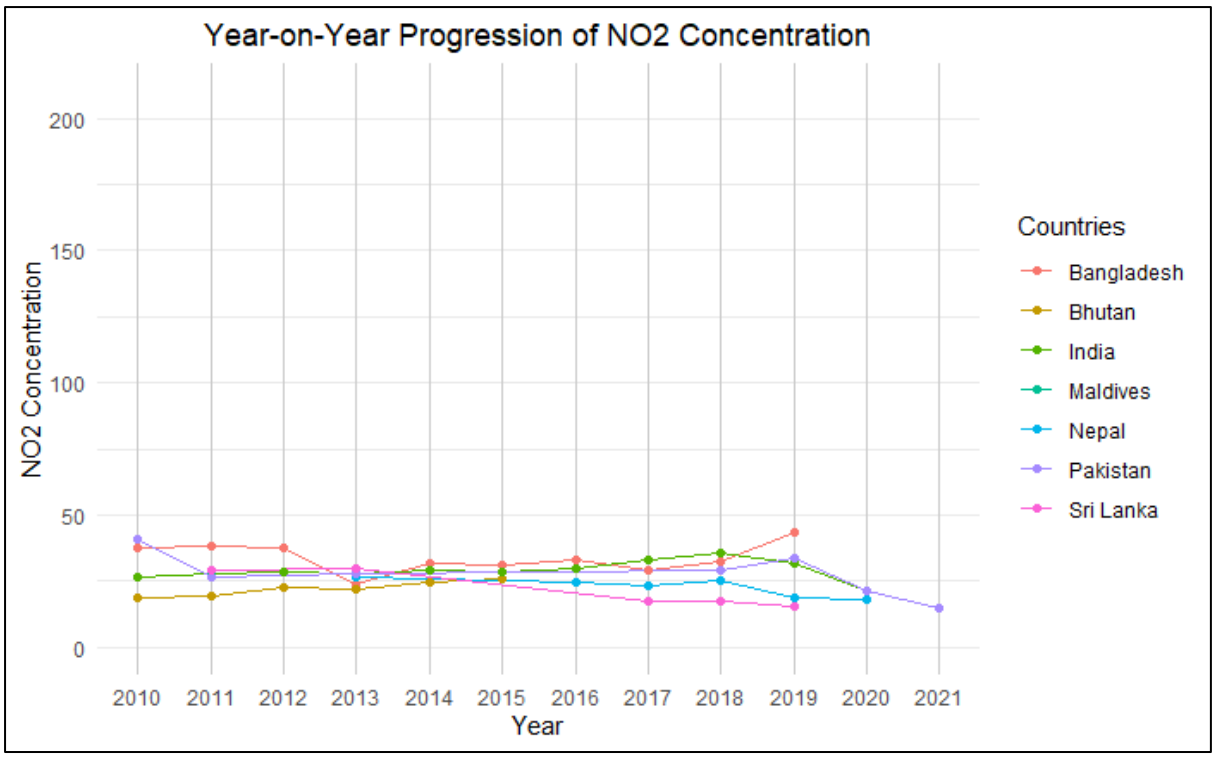

Fig B.18. Year-on-year NO_2 Progression by country.

Plots for section '4.2.3. Year-on-year Progression on concern level'.

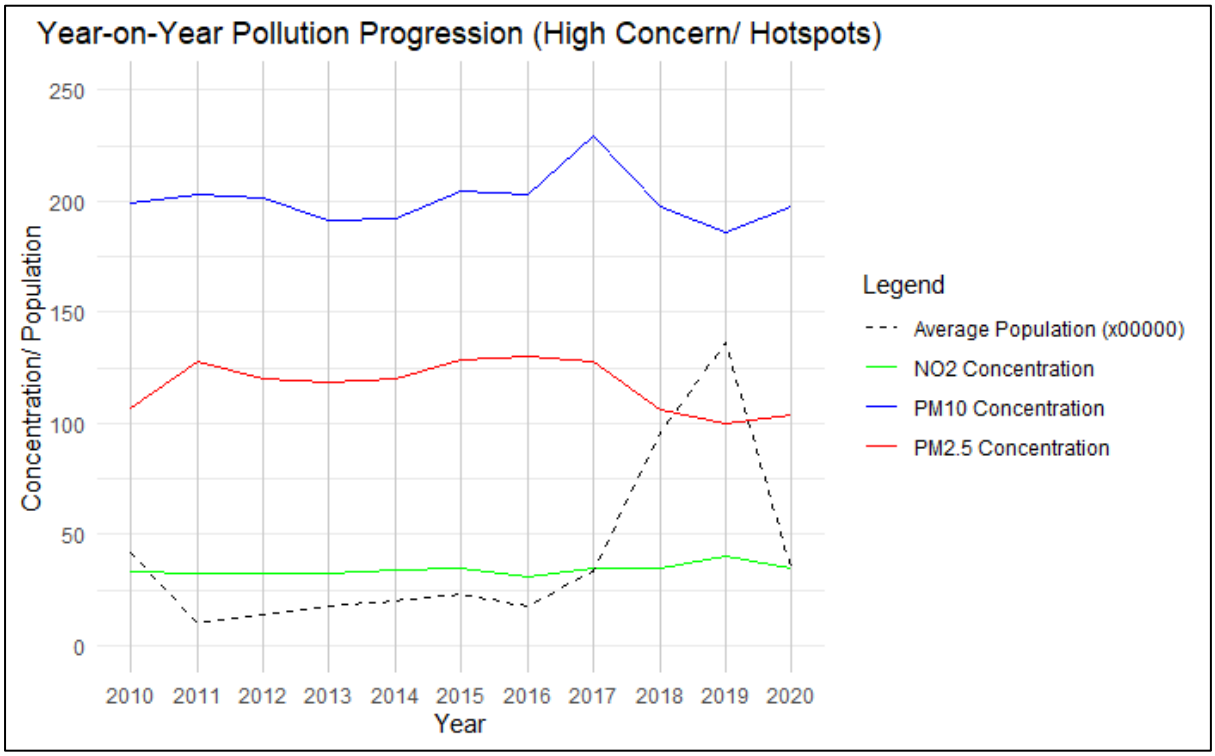

Fig B.19. Year-on-year Pollution Progression in the High concern cities or the Hotspots.

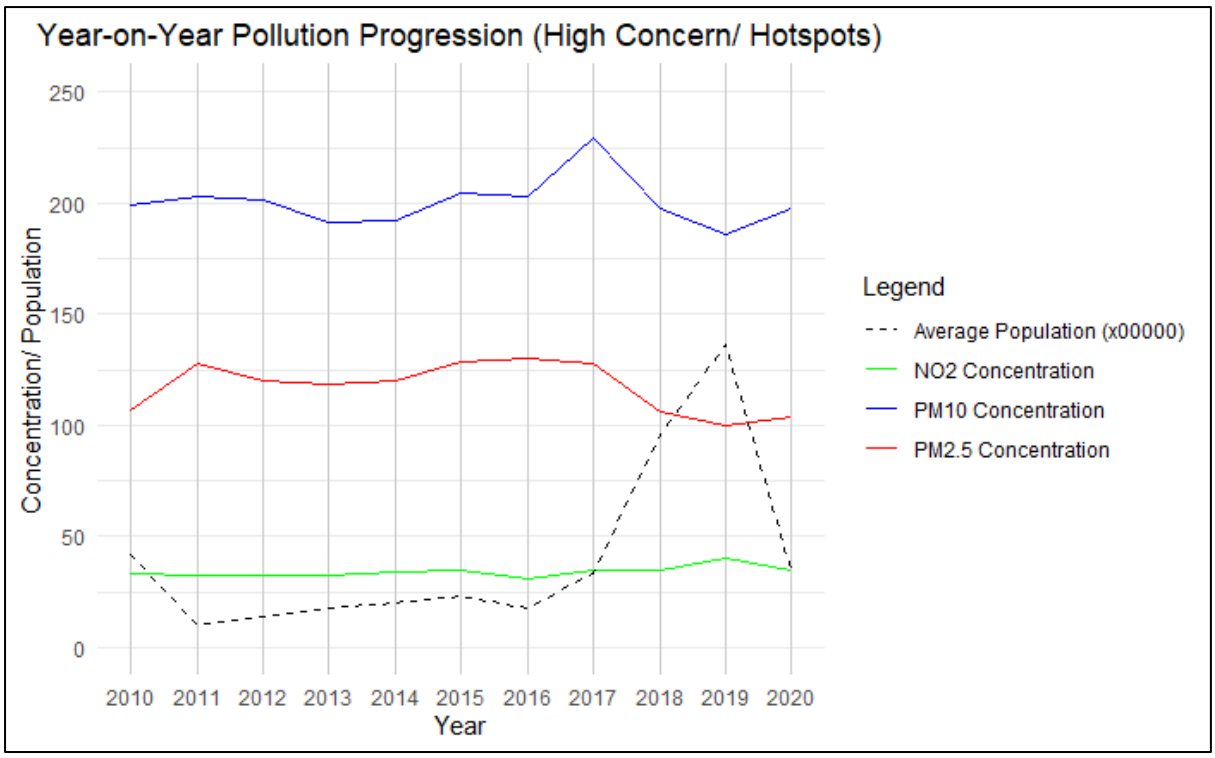

Fig B.20. Year-on-year Pollution Progression in the Medium concern cities.

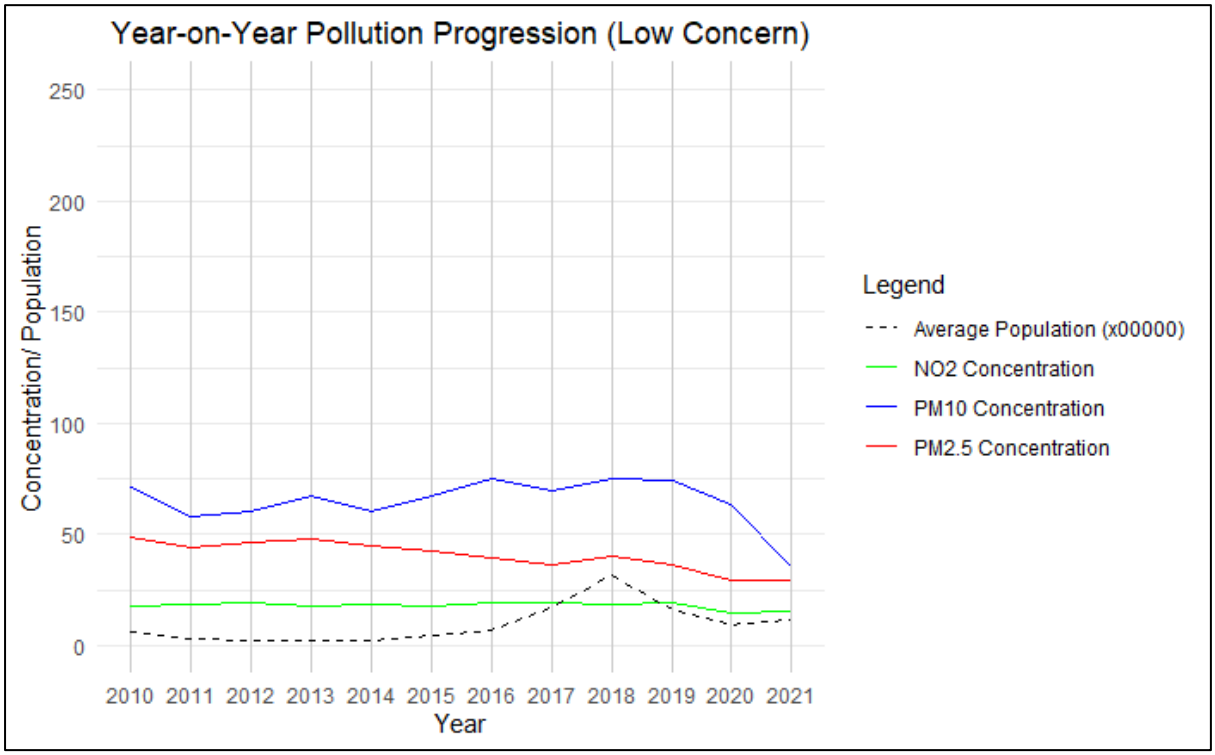

Fig B.21. Year-on-year Pollution Progression in the Low concern cities.

Appendix VI.

(i) Plots for ‘Relation with the 'Growth of urban areas' data’ sub-section.

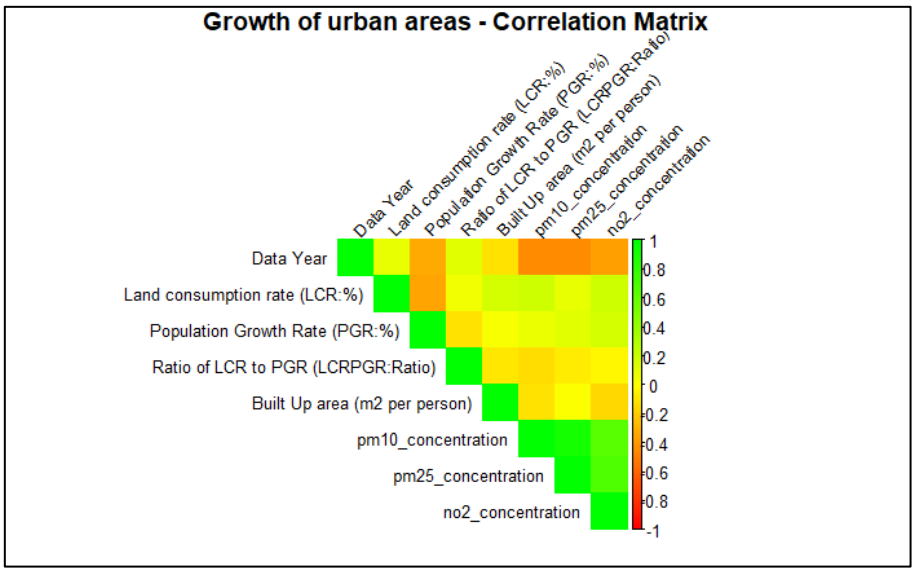

Fig B.22. Visual representation of correlation matrix of the ‘Growth of urban areas’ dataset.

(a) Regression Analysis - PM10 levels

Residuals:

Min	1Q	Median	3Q	Max
-70.30	-30.75	-12.22	17.04	149.93

Coefficients:	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
Intercept	94.24	31.34	3.01	0.006
LCR	10.24	7.82	1.31	0.203
PGR	7.81	9.12	0.86	0.400
Built Up area	-0.42	0.53	-0.80	0.429

Residual standard error: 55.44 on 25 degrees of freedom

Multiple R-squared: 0.0792 Adjusted R-squared: -0.0313

F-statistic: 0.7167 on 3 and 25 DF p-value: 0.5513

Table B.6. Summary of multiple regression analysis of PM10 on ‘Growth of urban areas’ data.

	2.5%	**97.5%**
Intercept	29.69	158.78
LCR	-5.87	26.34
PGR	-10.98	26.60
Built Up area	-1.51	0.66

Table B.7. Confidence intervals of multiple regression analysis of PM10 on ‘Growth of urban areas’ data.

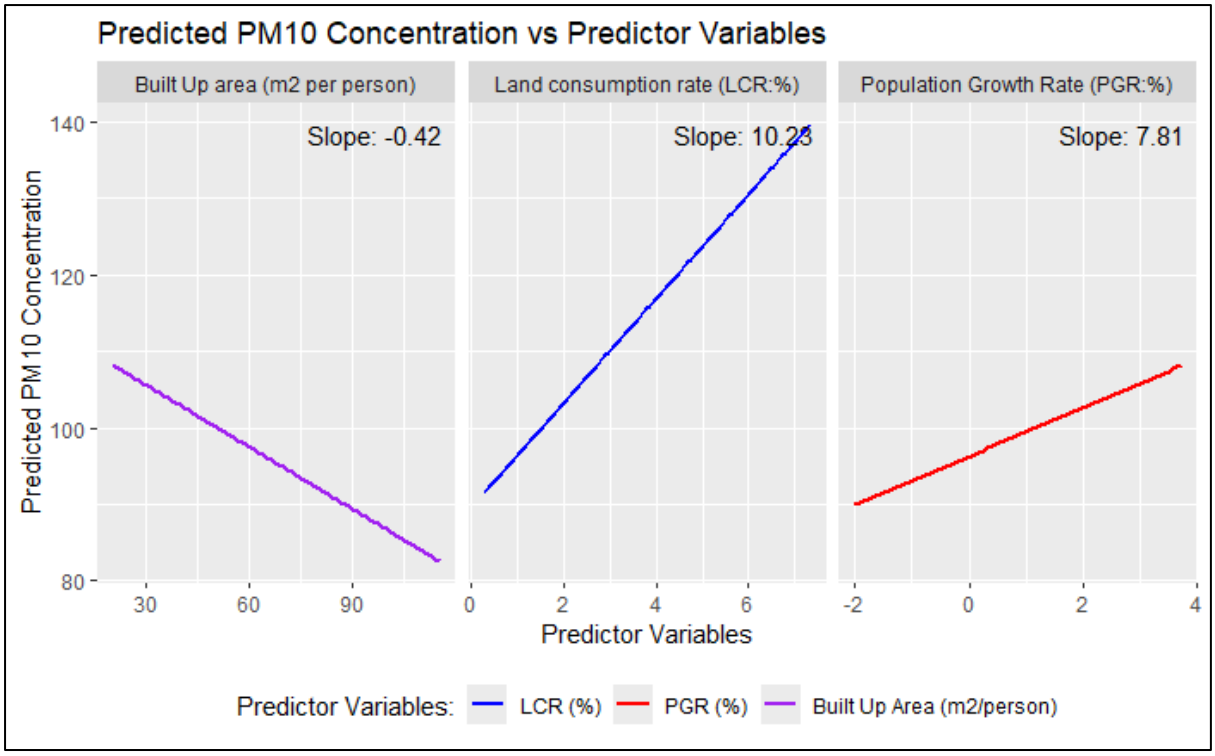

Fig B.23. Visual representation of regression analyses made for PM10 on the 'Growth of urban areas' dataset.

(b) Regression Analysis – PM2.5 levels

Residuals:

Min	1Q	Median	3Q	Max
-33.374	-16.924	-7.161	10.909	88.946

Coefficients:	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
Intercept	41.09	16.16	2.54	0.018
LCR	2.93	4.03	0.73	0.474
PGR	3.86	4.70	0.82	0.420
Built Up area	-0.03	0.27	-0.10	0.918

Residual standard error: 28.58 on 25 degrees of freedom

Multiple R-squared: 0.0342 Adjusted R-squared: -0.0818

F-statistic: 0.2946 on 3 and 25 DF p-value: 0.8289

Table B.8. Summary of multiple regression analysis of PM2.5 on 'Growth of urban areas' data.

	2.5%	**97.5%**
Intercept	7.81	74.37
LCR	-5.37	11.24
PGR	-5.83	13.54
Built Up area	-0.59	0.53

Table B.9. Confidence intervals of multiple regression analysis of PM2.5 on 'Growth of urban areas' data.

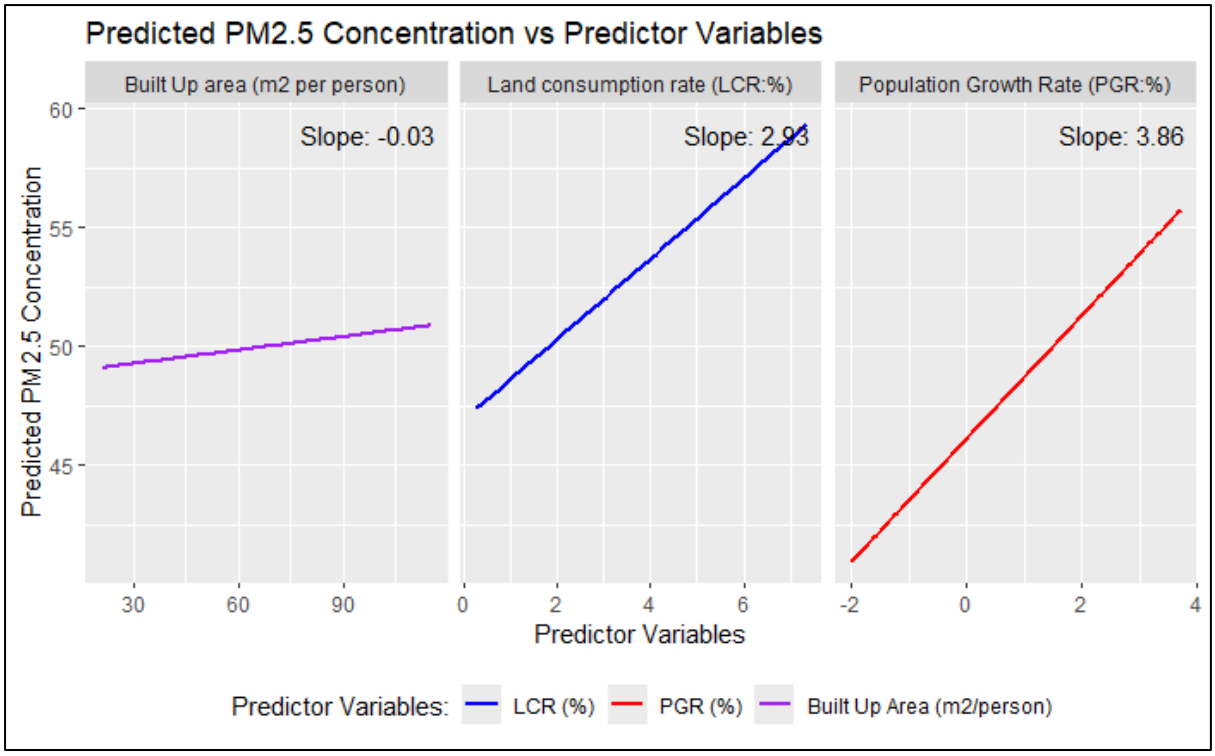

Fig B.24. Visual representation of regression analyses made for PM2.5 on the 'Growth of urban areas' dataset.

(c) Regression Analysis – NO_2 levels

Residuals:

Min	1Q	Median	3Q	Max
-20.498	-8.243	-4.263	10.619	32.007

Coefficients:	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
Intercept	24.09	7.09	3.40	0.002
LCR	2.90	1.77	1.64	0.114
PGR	3.05	2.06	1.48	0.152
Built Up area	-0.14	0.12	-1.21	0.237

Residual standard error: 12.55 on 25 degrees of freedom

Multiple R-squared: 0.1464 Adjusted R-squared: 0.0439

F-statistic: 1.429 on 3 and 25 DF p-value: 0.2579

Table B.10. Summary of multiple regression analysis of NO_2 on 'Growth of urban areas' data.

	2.5%	**97.5%**
Intercept	9.48	38.70
LCR	-0.75	6.54
PGR	-1.20	7.30
Built Up area	-0.39	0.10

Table B.11. Confidence intervals of multiple regression analysis of NO_2 on 'Growth of urban areas' data.

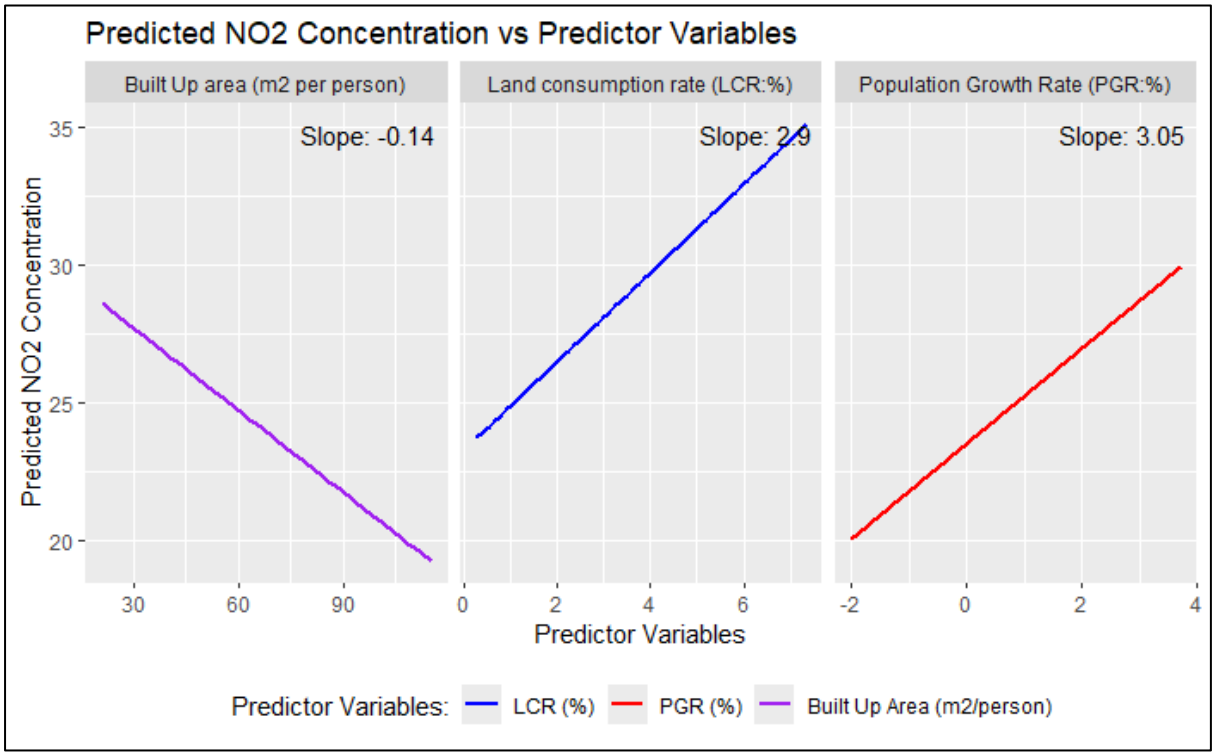

Fig B.25. Visual representation of regression analyses made for NO_2 on the 'Growth of urban areas' dataset.

(ii) Plots for 'Relation with the 'Urban transport' data' sub-section.

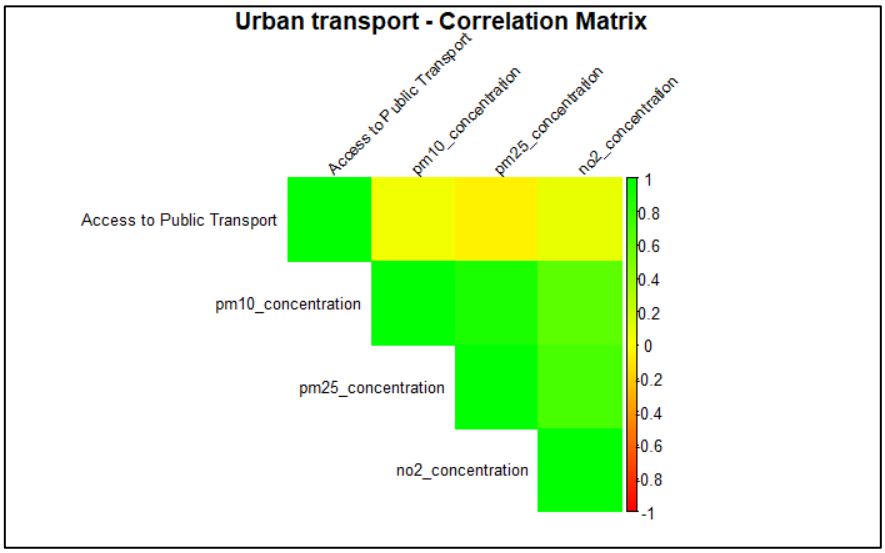

Fig B.26. Visual representation of correlation matrix of the 'Urban transport' dataset.

(a) Regression Analysis - PM10 levels

Residuals:				
Min	1Q	Median	3Q	Max
-52.34	-31.40	-17.57	21.16	102.42
Coefficients:	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
Intercept	85.43	18.13	4.71	0.0001
Access to Public Transport	-0.10	0.38	0.28	0.786
Residual standard error:	44.83 on 22 degrees of freedom			
Multiple R-squared:	0.0034		Adjusted R-squared:	-0.0429
F-statistic:	0.07556 on 1 and 22 DF		p-value:	0.786

Table B.12. Summary of regression analysis of PM10 on 'Urban transport' data.

	2.5%	**97.5%**
Intercept	47.82	123.04
Access to Public Transport	-0.68	0.89

Table B.13. Confidence intervals of regression analysis of PM10 on 'Urban transport" data.

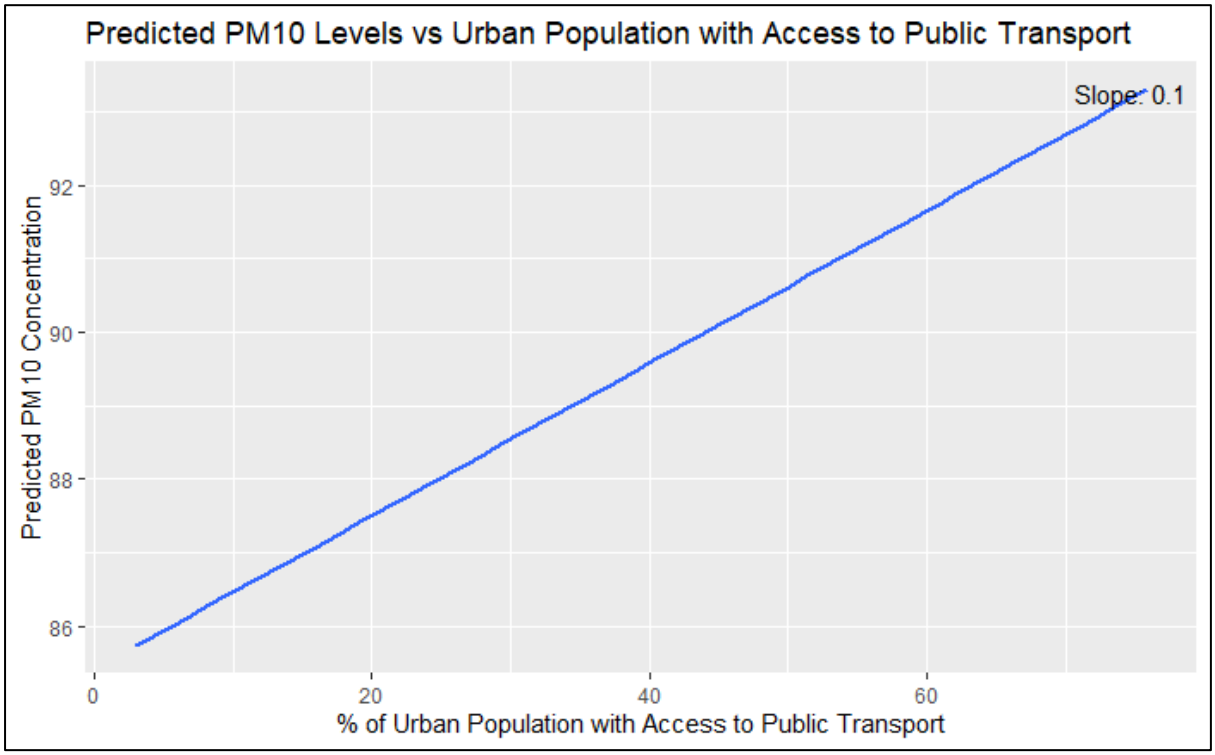

Fig B.27. Visual representation of regression analyses made for PM10 on the 'Urban transport' dataset.

(b) Regression Analysis – PM2.5 levels

Residuals:				
Min	1Q	Median	3Q	Max
-22.707	-14.473	-7.158	12.205	70.759
Coefficients:	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
Intercept	46.34	9.34	4.96	5.77e-05
Access to Public Transport	-0.04	0.19	-0.22	0.832
Residual standard error:	23.08 on 22 degrees of freedom			
Multiple R-squared:	0.0021		Adjusted R-squared:	-0.0433
F-statistic:	0.0463 on 1 and 22 DF		p-value:	0.8316

Table B.14. Summary of regression analysis of PM2.5 on 'Urban transport' data.

	2.5%	**97.5%**
Intercept	26.97	65.70
Access to Public Transport	-0.45	0.36

Table B.15. Confidence intervals of regression analysis of PM2.5 on 'Urban transport' data.

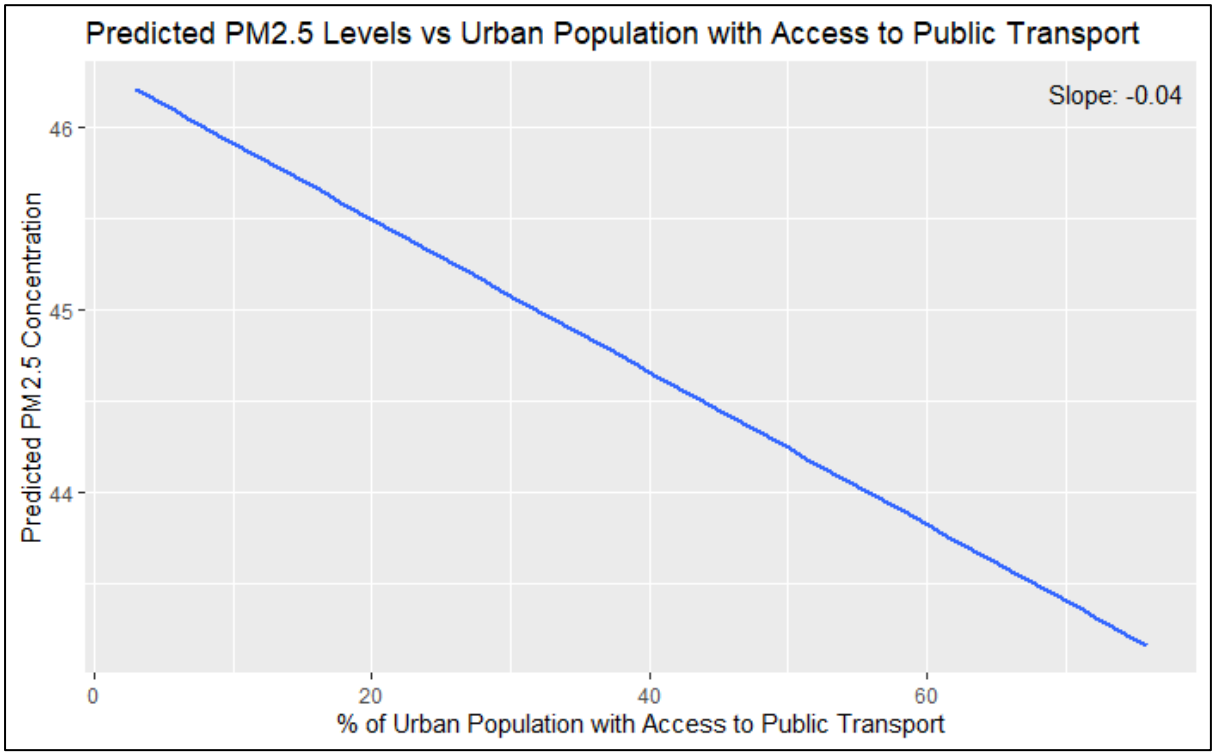

Fig B.28. Visual representation of regression analyses made for PM2.5 on the 'Urban transport' dataset.

(c) Regression Analysis – NO_2 levels

Residuals:				
Min	1Q	Median	3Q	Max
-15.164	-8.077	-4.958	7.599	36.452
Coefficients:	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
Intercept	21.63	5.03	4.30	0.0003
Access to Public Transport	0.04	0.10	0.43	0.673
Residual standard error:	12.43 on 22 degrees of freedom			
Multiple R-squared:	0.0082		Adjusted R-squared:	-0.0368
F-statistic:	0.1828 on 1 and 22 DF		p-value:	0.6732

Table B.16. Summary of regression analysis of PM2.5 on 'Urban transport' data.

	2.5%	**97.5%**
Intercept	11.20	32.05
Access to Public Transport	-0.17	0.26

Table B.17. Confidence intervals of regression analysis of PM2.5 on 'Urban transport' data.

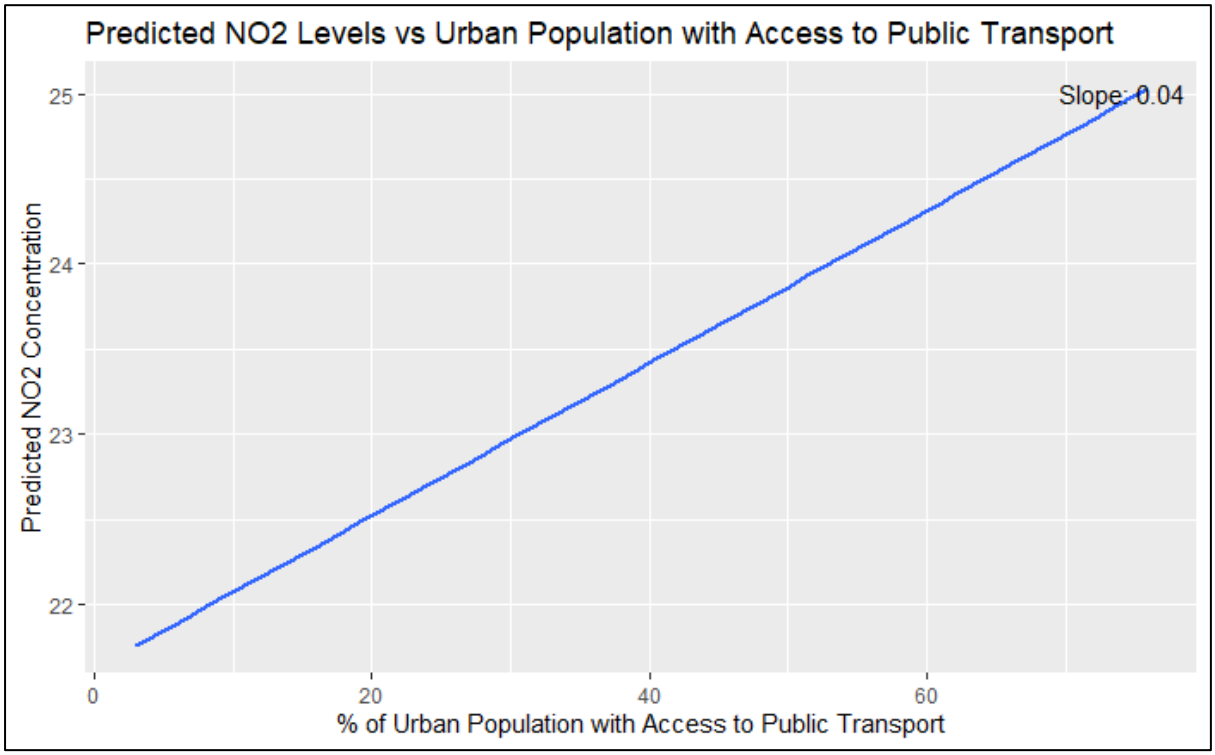

Fig B.29. Visual representation of regression analyses made for NO_2 on the 'Urban transport' dataset.

(iii) Plots for ‘Relation with the 'Open & Green Spaces' data’ sub-section.

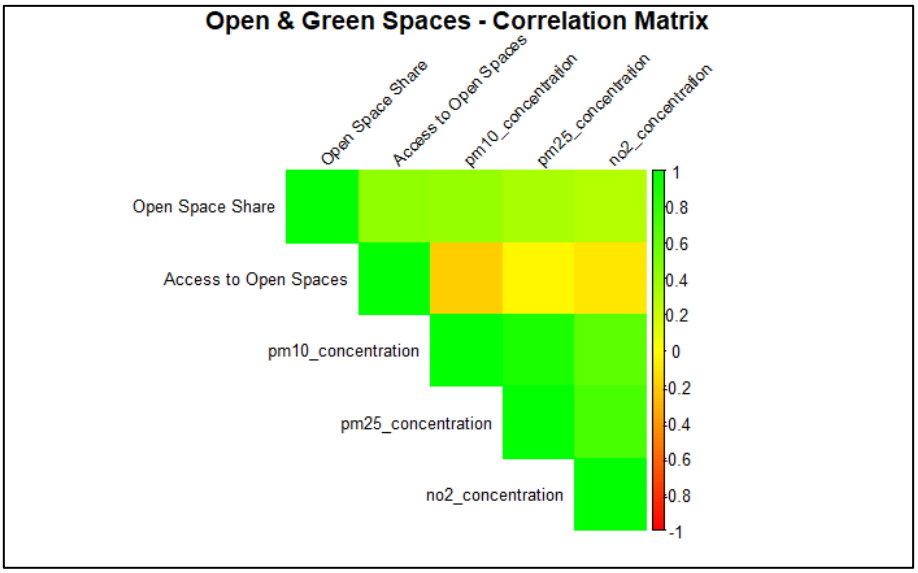

Fig B.30. Visual representation of correlation matrix of the 'Open & Green Spaces’ dataset.

(a) Regression Analysis - PM10 levels

Residuals:

Min	1Q	Median	3Q	Max
-69.609	-19.911	-1.137	29.118	68.940

Coefficients:	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
Intercept	39.57	36.13	1.10	0.287
Open Space Share	6.55	2.14	3.07	0.006
Access to Open Spaces	-1.67	0.70	-2.38	0.028

Residual standard error: 37.86 on 20 degrees of freedom

Multiple R-squared: 0.3493 Adjusted R-squared: 0.2842

F-statistic: 5.368 on 2 and 20 DF p-value: 0.0136

Table B.18. Summary of multiple regression analysis of PM10 on ‘Open & Green Spaces’ data.

	2.5%	**97.5%**
Intercept	-35.81	114.94
Open Space Share	2.09	11.00
Access to Open Spaces	-3.14	-0.20

Table B.19. Confidence intervals of multiple regression analysis of PM10 on ‘Open & Green Spaces’ data.

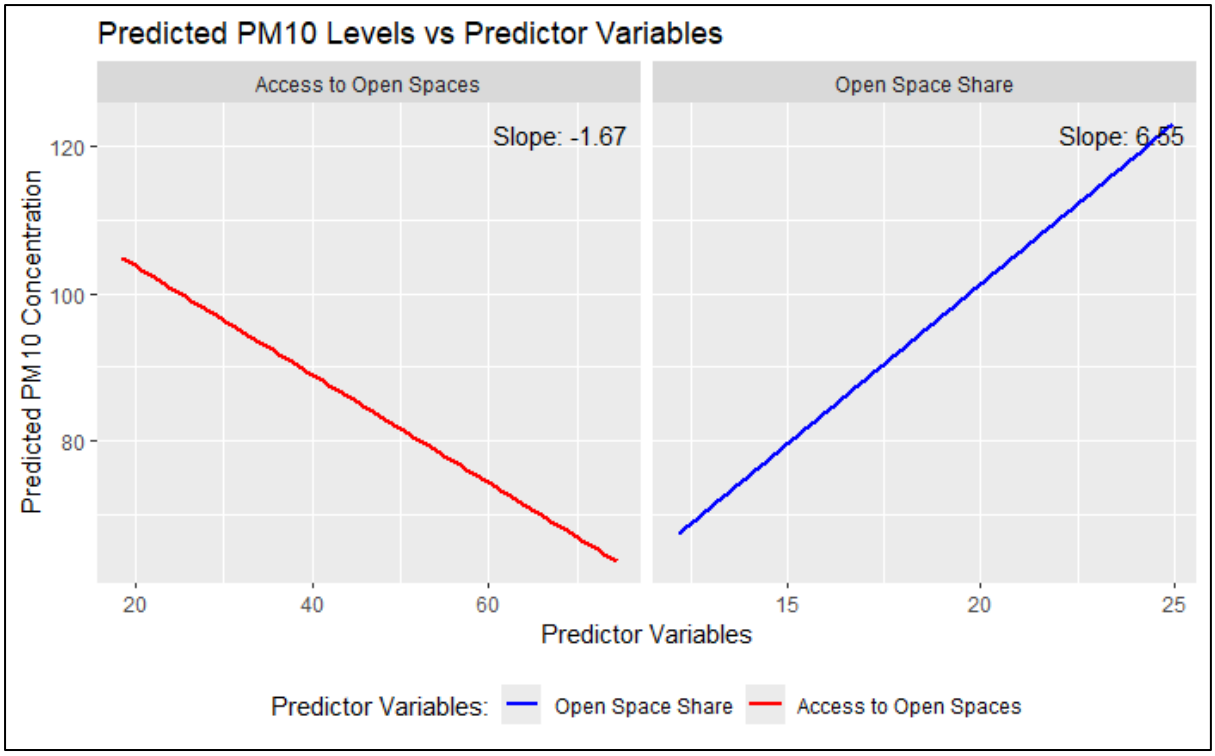

Fig B.31. Visual representation of regression analyses made for PM10 on the 'Open & Green Spaces' dataset.

(b) Regression Analysis – PM2.5 levels

Residuals:

Min	1Q	Median	3Q	Max
-37.280	-12.093	-3.573	12.577	56.010

Coefficients:	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
Intercept	19.32	21.29	0.91	0.375
Open Space Share	2.35	1.26	1.87	0.077
Access to Open Spaces	-0.41	0.42	-0.98	0.338

Residual standard error: 22.3 on 20 degrees of freedom

Multiple R-squared: 0.1496 Adjusted R-squared: 0.0646

F-statistic: 1.76 on 2 and 20 DF p-value: 0.1977

Table B.20. Summary of multiple regression analysis of PM2.5 on 'Open & Green Spaces' data.

	2.5%	**97.5%**
Intercept	-25.08	63.73
Open Space Share	-0.28	4.97
Access to Open Spaces	-1.27	0.46

Table B.21. Confidence intervals of multiple regression analysis of PM2.5 on 'Open & Green Spaces' data.

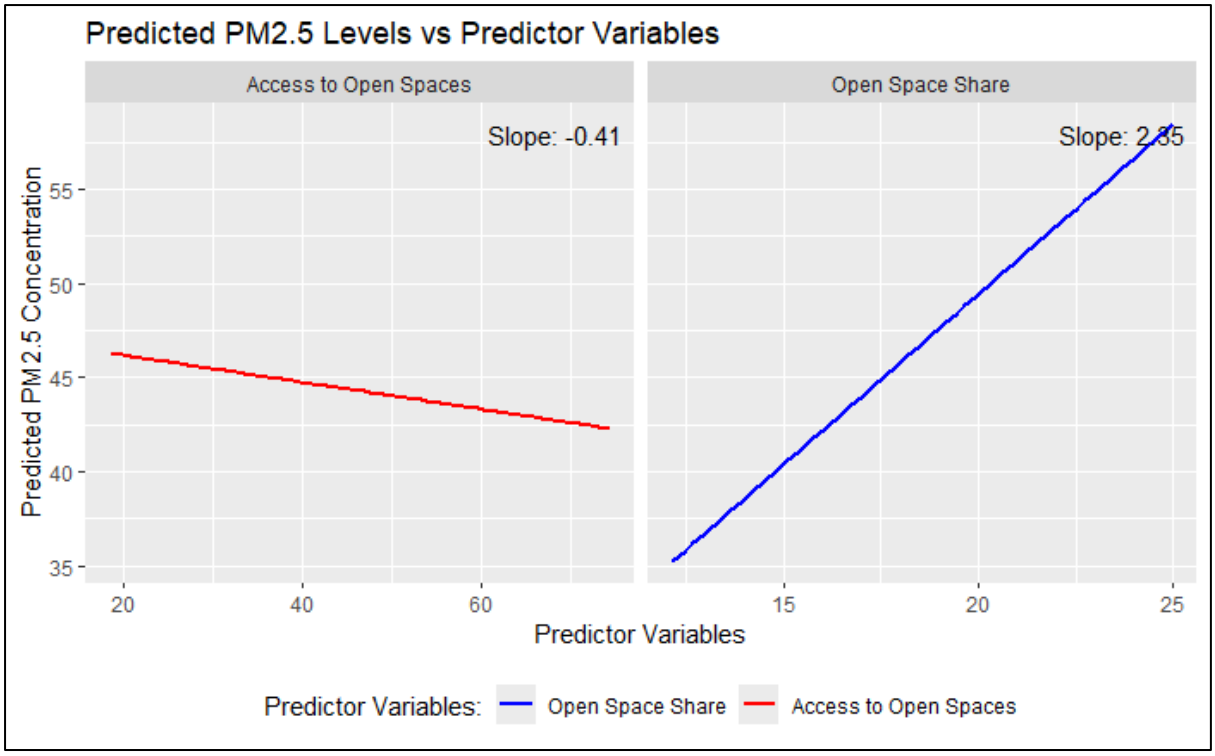

Fig B.32. Visual representation of regression analyses made for PM2.5 on the 'Open & Green Spaces' dataset.

(c) Regression Analysis – NO_2 levels

Residuals:

Min	1Q	Median	3Q	Max
-14.752	-7.526	-3.547	3.192	30.359

Coefficients:	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
Intercept	12.98	11.43	1.14	0.269
Open Space Share	1.23	0.68	1.82	0.083
Access to Open Spaces	-0.28	0.22	-1.27	0.219

Residual standard error: 11.97 on 20 degrees of freedom

Multiple R-squared: 0.1527 Adjusted R-squared: 0.0679

F-statistic: 1.802 on 2 and 20 DF p-value: 0.1908

Table B.22. Summary of multiple regression analysis of NO_2 on 'Open & Green Spaces' data.

	2.5%	**97.5%**
Intercept	-10.85	36.82
Open Space Share	-0.18	2.64
Access to Open Spaces	-0.75	0.18

Table B.23. Confidence intervals of multiple regression analysis of NO_2 on 'Open & Green Spaces' data.

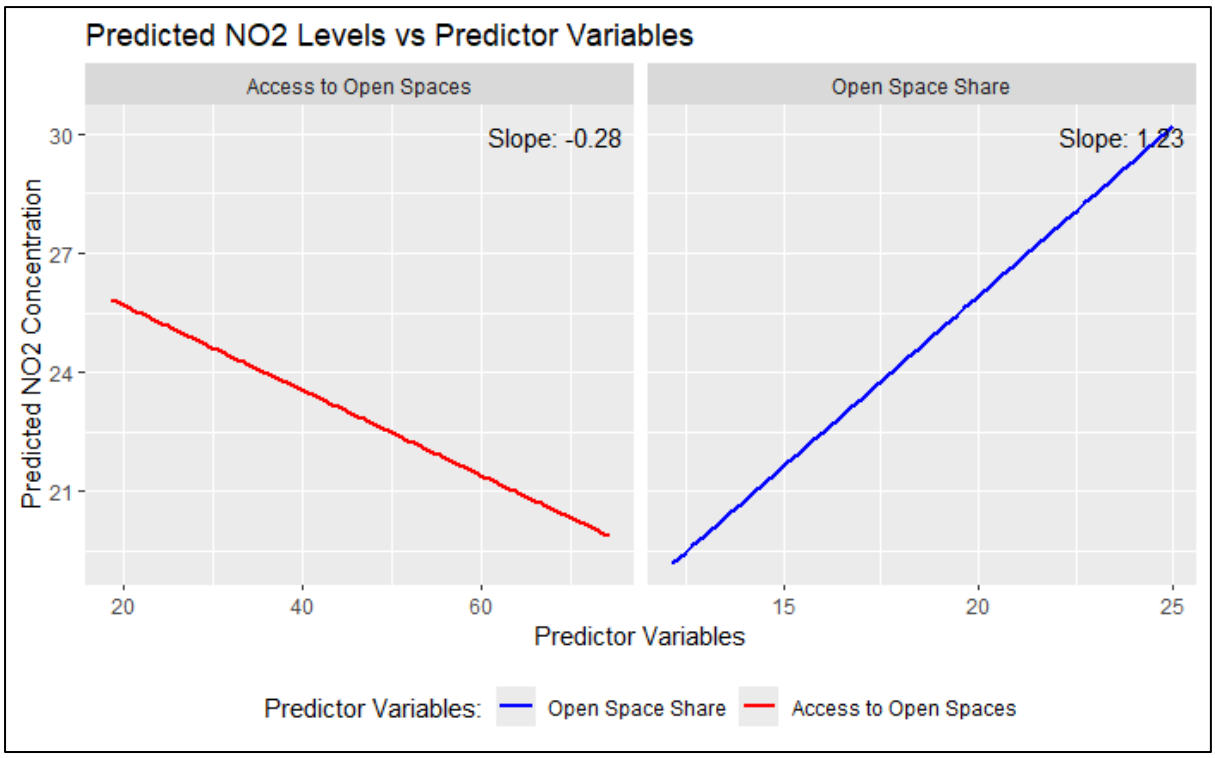

Fig B.33. Visual representation of regression analyses made for NO_2 on the 'Open & Green Spaces' dataset.

(iv) Plots for 'Relation with the Population data' sub-section.

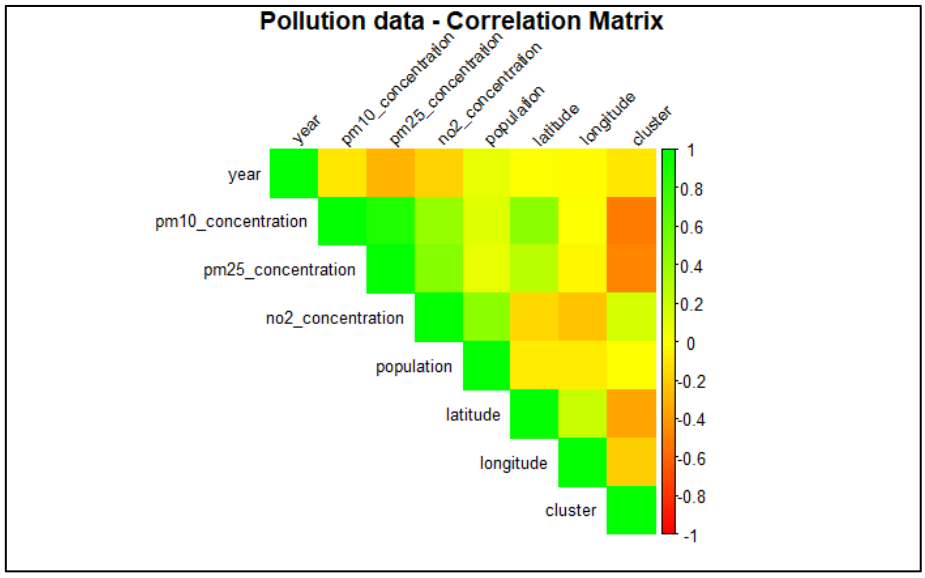

Fig B.34. Visual representation of correlation matrix of the 'Pollution dataset.

(a) Regression Analysis - PM10 levels

Residuals:				
Min	1Q	Median	3Q	Max
-90.61	-39.14	-12.61	27.66	208.16
Coefficients:	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
Intercept	1.000e+02	1.582e+00	63.231	<2e-16
population	1.903e-06	3.874e-07	4.914	1e-06
Residual standard error:	53.57 on 1381 degrees of freedom			
Multiple R-squared:	0.0172		Adjusted R-squared:	0.0165
F-statistic:	24.14 on 1 and 1381 DF		p-value:	1.001e-06

Table B.24. Summary of regression analysis of PM10 on 'Pollution' data.

	2.5%	**97.5%**
Intercept	9.690432e+01	1.031096e+02
population	1.143426e-06	2.663193e-06

Table B.25. Confidence intervals of regression analysis of PM10 on 'Pollution' data.

(b) Regression Analysis – PM2.5 levels

Residuals:				
Min	1Q	Median	3Q	Max
-54.159	-17.460	-6.689	8.240	117.082
Coefficients:	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
Intercept	6.406e+01	8.900e-01	71.972	<2e-16
population	6.940e-07	2.180e-07	3.184	0.0015
Residual standard error:	30.15 on 1381 degrees of freedom			
Multiple R-squared:	0.0073		Adjusted R-squared:	0.0066
F-statistic:	10.13 on 1 and 1381 DF		p-value:	0.0015

Table B.26. Summary of regression analysis of PM2.5 on 'Pollution' data.

	2.5%	**97.5%**
Intercept	6.231218e+01	6.580415e+01
population	2.663389e-07	1.121576e-06

Table B.27. Confidence intervals of regression analysis of PM2.5 on 'Pollution' data.

(c) Regression Analysis – NO_2 levels

Residuals:				
Min	1Q	Median	3Q	Max
-24.7528	-4.7616	0.4683	5.3053	28.1086

Coefficients:	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
Intercept	2.588e+01	2.369e-01	109.26	<2e-16
population	1.089e-06	5.801e-08	18.78	<2e-16

Residual standard error: 8.023 on 1381 degrees of freedom

Multiple R-squared: 0.2034 Adjusted R-squared: 0.2028

F-statistic: 352.6 on 1 and 1381 DF p-value: <2.2e-16

Table B.28. Summary of regression analysis of NO_2 on 'Pollution' data.

	2.5%	**97.5%**
Intercept	2.541532e+01	2.634462e+01
population	9.754654e-07	1.203065e-06

Table B.29. Confidence intervals of regression analysis of NO_2 on 'Pollution' data.

Appendix VII.

```
# Importing the libraries
library(tidyverse)
library(ggplot2)
library(dplyr)
library(missForest)
library(cluster)
library(factoextra)
library(openxlsx)
library(Hmisc)
library(reshape2)
library(corrplot)
countries <- c('Afghanistan','Bangladesh','Bhutan','India','Maldives','Nepal','Pakistan','Sri Lanka')
```

Importing Raw Data

```
# Importing Pollution Data
pollution_rawdata <- readxl::read_excel(
'who_ambient_air_quality_database_version_2024_(v6.1).xlsx', sheet = "Update 2024 (V6.1)")
#Importing Socio-Economic Data
Open_and_green_spaces_rawdata <- readxl::read_excel('SDG_11-7-1.xlsx')
Growth_of_urban_areas_rawdata <- readxl::read_excel('SDG_11-3-1.xlsx')
Urban_transport_rawdata <- readxl::read_excel('SDG_11-2-1.xlsx')
```

Filtering the data to keep only South Asian countries

```
# Filter for South Asia
pollution_rawdata <- filter(pollution_rawdata,pollution_rawdata$country_name %in% countries)
Open_and_green_spaces_rawdata <- filter(Open_and_green_spaces_rawdata,
Open_and_green_spaces_rawdata$`Country or Territory Name` %in% countries)
Growth_of_urban_areas_rawdata <- filter(Growth_of_urban_areas_rawdata,
Growth_of_urban_areas_rawdata$`Country or Territory Name` %in% countries)
Urban_transport_rawdata <- filter(Urban_transport_rawdata, Urban_transport_rawdata$`Country or
Territory Name` %in% countries)
```

Pulling the data as tibbles

1.) Pollution Data

```
head(pollution_rawdata)
# Keeping only the essential fields
pollution_rawdata <- pollution_rawdata %>% select(country_name, city, year, pm10_concentration,
pm25_concentration, no2_concentration, population, latitude, longitude)
```

2.) Socio-Economic Data

2.1.) Open Spaces and Green Areas Data

```
head(Open_and_green_spaces_rawdata)
# Keeping only the essential fields
Open_and_green_spaces_rawdata <- Open_and_green_spaces_rawdata %>%
select(`Country or Territory Name`, `City Name`,
`Average share of the built-up area of cities that is open space for public use for all (%) [a]`,
`Average share of urban population with convenient access to open public spaces (%) [b]`,
`Data Reference Year`)
```

2.2.) Spatial growth of cities and urban areas Data

```
head(Growth_of_urban_areas_rawdata)
# Keeping only the essential fields
Growth_of_urban_areas_rawdata <- Growth_of_urban_areas_rawdata %>%
select(`Country or Territory Name`, `City Name`, `Data Year 1`, `Data Year 2`,
`Data Year 3`, `Land consumption rate (LCR) Year 1 to Year 2 (%)`,
`Land consumption rate (LCR) Year 2 to Year 3 (%)`,
`Population Growth Rate (PGR) Year 1 to Year 2 (%)`,
`Population Growth Rate (PGR) Year 2 to Year 3 (%)`,
`Ratio of Land Consumption Rate to Population Growth Rate (LCRPGR) Year 1 to Year 2 (Ratio)`,
`Ratio of Land Consumption Rate to Population Growth Rate (LCRPGR) Year 2 to Year 3 (Ratio)`,
`Built Up area Per Capita Year 1 (m2 per person)`,
`Built Up area Per Capita Year 2 (m2 per person)`,
`Built Up area Per Capita Year 3 (m2 per person)`)
```

2.3.) Urban Transport Data

```
head(Urban_transport_rawdata)
# Keeping only the essential fields
Urban_transport_rawdata <- Urban_transport_rawdata %>%
select(`Country or Territory Name`, `City Name`,
`Share of urban population with convenient access to public transport (%)`,
`Data Reference Year`)
```

Pulling data summary of each dataset.

```
summary(pollution_rawdata)
summary(Open_and_green_spaces_rawdata)
summary(Growth_of_urban_areas_rawdata)
summary(Urban_transport_rawdata)
```

A couple of issues noted above:
1.) For the consistency of names of cities, we need to update the 'Pollution' dataset, since the city names also include country code in them which is not the case with other datasets.

```
pollution_rawdata$city <- sub("/.*", "", pollution_rawdata$city)
```

2.) Variables like 'pm10_concentration', 'pm25_concentration', 'no2_concentration', 'population' in the 'Pollution' dataset are not imported as numeric, at the same time 'country_name' and 'city' are not imported as factor, and need to be converted accordingly.

```
columns_to_convert <- c("pm10_concentration", "pm25_concentration", "no2_concentration", "population")
pollution_rawdata[columns_to_convert] <- lapply(pollution_rawdata[columns_to_convert], as.numeric)
columns_to_factor <- c("country_name","city")
pollution_rawdata[columns_to_factor] <- lapply(pollution_rawdata[columns_to_factor], as.factor)
```

Identifying missing values

```
no_of_missing_values <- data.frame(datasets=c('Urban transport', 'Open & Green Spaces', 'Growth of urban areas', 'Pollution'),
missing_values_proportion=missing_values_proportion=c((sum(is.na(Urban_transport_rawdata))/nrow(Urban_transport_rawdata)*ncol(Urban_transport_rawdata)),(sum(is.na(Open_and_green_spaces_rawdata))/nrow(Open_and_green_spaces_rawdata)*ncol(Open_and_green_spaces_rawdata)),(sum(is.na(Growth_of_urban_areas_rawdata))/nrow(Growth_of_urban_areas_rawdata)*ncol(Growth_of_urban_areas_rawdata)),(sum(is.na(pollution_rawdata))/nrow(pollution_rawdata)*ncol(pollution_rawdata))))

no_of_missing_values
```

To deal with the missing values we first need to understand what type of data is missing.

```
summary(Urban_transport_rawdata)
summary(Open_and_green_spaces_rawdata)
summary(Growth_of_urban_areas_rawdata)
summary(pollution_rawdata)
```

Here is how the above missing values can be dealt with:

- Open & Green Spaces data: All the 22 missing values are in the 'Average share of the built-up area of cities that is open space for public use for all (%) [a]' field, therefore, for these 22 records we can skip this field.
- Growth of urban areas data: All the 24 missing values are in the 'Ratio of Land Consumption Rate to Population Growth Rate (LCRPGR) Year 2 to Year 3 (Ratio)' field, which is basically a ratio of 'Land consumption rate (LCR) Year 2 to Year 3 (%)' and 'Population Growth Rate (PGR) Year 2 to Year 3 (%)'. Therefore, we can simply calculate all the missing values for this dataset.
- Pollution data: 3 missing values are are in form of 'year', 55 in 'pm10_concentration', 49 in 'population', and ~1100 in 'pm25_concentration' and 'no2_concentration'. Since, 'year' is very essential for those particular records, therefore it is better to out rightly remove those 3 records. For all other missing values imputation can be used.

Checking if values are missing at random or not.

```
# For 'Open & Green Spaces' data
Open_and_green_spaces_NAdata <- filter(Open_and_green_spaces_rawdata, is.na(`Average share of the built-up area of cities that is open space for public use for all (%) [a]`))
summary(Open_and_green_spaces_NAdata)
```

Since, there are no clear differences in the summary of the 'Open & Green Spaces' data with and without the NAs. We can conclude that the values in 'Open & Green Spaces' data are Missing Completely at Random (MCAR). Since all the 22 missing values are in the 'Average share of the built-up area of cities that is open space for public use for all (%) [a]' variable, we can skip these 22 records for this field.

Now checking for the 'Pollution' dataset.

```
# For 'Pollution' data - 'year' column
pollution_yearNA <- filter(pollution_rawdata, is.na(year))
summary(pollution_yearNA)

# For 'Pollution' data - 'pm10_concentration' column
pollution_pm10NA <- filter(pollution_rawdata, is.na(pm10_concentration))
summary(pollution_pm10NA)

# For 'Pollution' data - 'pm25_concentration' column
pollution_pm25NA <- filter(pollution_rawdata, is.na(pm25_concentration))
summary(pollution_pm25NA)

# For 'Pollution' data - 'no2_concentration' column
pollution_no2NA <- filter(pollution_rawdata, is.na(no2_concentration))
summary(pollution_no2NA)

# For 'Pollution' data - 'population' column
pollution_populationNA <- filter(pollution_rawdata, is.na(population))
summary(pollution_populationNA)
```

Above summaries, clearly show that the missing values are spread across countries and cities. Only 2 exceptions are missing values in 'pm25_concentration' and 'no2_concentration', which are mostly from India, however they are still spread across cities within India. Therefore, the missing values in the 'Pollution' data also are MCAR. Since, 'year' is very essential for those particular records, therefore it is better to out rightly remove those 3 records. For all other missing values imputation can be used.

Fixing the data issues.

```
# Filling back the missing values in the Growth of urban areas data
Growth_of_urban_areas_rawdata <- Growth_of_urban_areas_rawdata %>%
mutate(`Ratio of Land Consumption Rate to Population Growth Rate (LCRPGR) Year 2 to Year 3 (Ratio)` =
ifelse(is.na(`Ratio of Land Consumption Rate to Population Growth Rate (LCRPGR) Year 2 to Year 3 (Ratio)`),
`Land consumption rate (LCR) Year 2 to Year 3 (%)` / `Population Growth Rate (PGR) Year 2 to Year 3 (%)`,
`Ratio of Land Consumption Rate to Population Growth Rate (LCRPGR) Year 2 to Year 3 (Ratio)`))

summary(Growth_of_urban_areas_rawdata)
```

```
# Removing 'NA' records in the 'year' column from Pollution data
pollution_rawdata <- pollution_rawdata[!is.na(pollution_rawdata$year), ]
summary(pollution_rawdata)
```

```
# Taking out factors before applying MissForest imputation
pollution_data_location <- pollution_rawdata[, c("country_name","city","year")]
pollution_rawdata <- as.data.frame(lapply(pollution_rawdata, function(x) {
if (is.character(x)) {
return(as.factor(x))
} else {
return(as.numeric(as.character(x)))
}
}))

# Applying MissForest imputation technique to deal with other missing values in the Pollution data
set.seed(10)
imputed_data <- missForest(pollution_rawdata)
pollution_imputed_data <- imputed_data$ximp
pollution_imputed_data <- pollution_imputed_data[, !(names(pollution_imputed_data) == "year")]
pollution_imputed_data <- cbind(pollution_data_location,pollution_imputed_data)

# Printing summary of the pollution data
summary(pollution_imputed_data)
```

Since 2021 dataset only has 1 record, we can skip this one record for rest of the analysis.

Checking the outliers

IQR Method to remove outliers from 'Open & Green Spaces', 'Urban transport', and 'Growth of urban areas' datasets

1. Removing outliers from 'Open & Green Spaces' dataset - not required since all numeric datapoints are in form of percentages.
2. Removing outliers from 'Urban transport' dataset - not required since all numeric datapoints are in form of percentages.
3. Removing outliers from 'Growth of urban areas' dataset –

```
# Checking outliers in `Built Up area Per Capita Year 1 (m2 per person)`
Q1_Urbangrowth_a <- quantile(Growth_of_urban_areas_rawdata$`Built Up area Per Capita Year 1 (m2 per person)`, 0.25, na.rm = TRUE)
Q3_Urbangrowth_a <- quantile(Growth_of_urban_areas_rawdata$`Built Up area Per Capita Year 1 (m2 per person)`, 0.75, na.rm = TRUE)
IQR_Urbangrowth_a <- Q3_Urbangrowth_a - Q1_Urbangrowth_a
outliers_Urbangrowth_a <- which(Growth_of_urban_areas_rawdata$`Built Up area Per Capita Year 1 (m2 per person)` < (Q1_Urbangrowth_a - 1.5 * IQR_Urbangrowth_a) |
            Growth_of_urban_areas_rawdata$`Built Up area Per Capita Year 1 (m2 per person)` >
(Q3_Urbangrowth_a + 1.5 * IQR_Urbangrowth_a))
```

```
# Checking outliers in `Built Up area Per Capita Year 2 (m2 per person)`
Q1_Urbangrowth_b <- quantile(Growth_of_urban_areas_rawdata$`Built Up area Per Capita Year 2 (m2 per person)`, 0.25, na.rm = TRUE)
Q3_Urbangrowth_b <- quantile(Growth_of_urban_areas_rawdata$`Built Up area Per Capita Year 2 (m2 per person)`, 0.75, na.rm = TRUE)
IQR_Urbangrowth_b <- Q3_Urbangrowth_b - Q1_Urbangrowth_b
outliers_Urbangrowth_b <- which(Growth_of_urban_areas_rawdata$`Built Up area Per Capita Year 2 (m2 per person)` < (Q1_Urbangrowth_b - 1.5 * IQR_Urbangrowth_b) |
            Growth_of_urban_areas_rawdata$`Built Up area Per Capita Year 2 (m2 per person)` >
(Q3_Urbangrowth_b + 1.5 * IQR_Urbangrowth_b))

# Checking outliers in `Built Up area Per Capita Year 3 (m2 per person)`
Q1_Urbangrowth_c <- quantile(Growth_of_urban_areas_rawdata$`Built Up area Per Capita Year 3 (m2 per person)`, 0.25, na.rm = TRUE)
Q3_Urbangrowth_c <- quantile(Growth_of_urban_areas_rawdata$`Built Up area Per Capita Year 3 (m2 per person)`, 0.75, na.rm = TRUE)
IQR_Urbangrowth_c <- Q3_Urbangrowth_c - Q1_Urbangrowth_c
outliers_Urbangrowth_c <- which(Growth_of_urban_areas_rawdata$`Built Up area Per Capita Year 3 (m2 per person)` < (Q1_Urbangrowth_c - 1.5 * IQR_Urbangrowth_c) |
            Growth_of_urban_areas_rawdata$`Built Up area Per Capita Year 3 (m2 per person)` >
(Q3_Urbangrowth_c + 1.5 * IQR_Urbangrowth_c))

# Combine the three sets of outliers
all_outliers_Urbangrowth <- unique(c(outliers_Urbangrowth_a, outliers_Urbangrowth_b,
outliers_Urbangrowth_c))

# Remove the outliers from the dataset
Growth_of_urban_areas_rawdata_wo_outliers <- Growth_of_urban_areas_rawdata[-
all_outliers_Urbangrowth, ]

summary(Growth_of_urban_areas_rawdata_wo_outliers)
```

```
# Comparing means of the datasets with and without outliers
comparing_means_GUA_datasets <- data.frame(datasets=c('With Outlier Capping', 'Without Outlier Capping'),
'Mean Built Up area Per Capita Year 1'= c(mean(Growth_of_urban_areas_rawdata$`Built Up area Per Capita Year 1 (m2 per person)`),mean(Growth_of_urban_areas_rawdata_wo_outliers$`Built Up area Per Capita Year 1 (m2 per person)`)), 'Mean Built Up area Per Capita Year 2'= c(mean (Growth_of_urban_areas_rawdata$`Built Up area Per Capita Year 2 (m2 per person)`), mean(Growth_of_urban_areas_rawdata_wo_outliers$`Built Up area Per Capita Year 2 (m2 per person)`)), 'Mean Built Up area Per Capita Year 3'=
c(mean(Growth_of_urban_areas_rawdata$`Built Up area Per Capita Year 3 (m2 per person)`),mean(Growth_of_urban_areas_rawdata_wo_outliers$`Built Up area Per Capita Year 3 (m2 per person)`)))

comparing_means_GUA_datasets
```

```
# Comparing medians of the datasets with and without outliers
comparing_medians_GUA_datasets <- data.frame(datasets=c('With Outlier Capping', 'Without Outlier Capping'), 'Median Built Up area Per Capita Year 1'= c(median(Growth_of_urban_areas_rawdata$`Built Up area Per Capita Year 1 (m2 per person)`), median(Growth_of_urban_areas_rawdata_wo_outliers$`Built Up area Per Capita Year 1 (m2 per person)`)), 'Median Built Up area Per Capita Year 2'=
c(median(Growth_of_urban_areas_rawdata$`Built Up area Per Capita Year 2 (m2 per person)`),
median(Growth_of_urban_areas_rawdata_wo_outliers$`Built Up area Per Capita Year 2 (m2 per person)`)),
'Median Built Up area Per Capita Year 3'= c(median(Growth_of_urban_areas_rawdata$`Built Up area Per Capita Year 3 (m2 per person)`), median(Growth_of_urban_areas_rawdata_wo_outliers$`Built Up area Per Capita Year 3 (m2 per person)`)))

comparing_medians_GUA_datasets
```

Clearly both the datasets have very different mean and median values.

```
# T-test on 'Built Up area Per Capita Year 1 (m2 per person)'
t.test(Growth_of_urban_areas_rawdata$`Built Up area Per Capita Year 1 (m2 per person)`,
Growth_of_urban_areas_rawdata_wo_outliers$`Built Up area Per Capita Year 1 (m2 per person)`)

# T-test on 'Built Up area Per Capita Year 2 (m2 per person)'
t.test(Growth_of_urban_areas_rawdata$`Built Up area Per Capita Year 2 (m2 per person)`,
Growth_of_urban_areas_rawdata_wo_outliers$`Built Up area Per Capita Year 2 (m2 per person)`)

# T-test on 'Built Up area Per Capita Year 3 (m2 per person)'
t.test(Growth_of_urban_areas_rawdata$`Built Up area Per Capita Year 3 (m2 per person)`,
Growth_of_urban_areas_rawdata_wo_outliers$`Built Up area Per Capita Year 3 (m2 per person)`)
```

Since mean values of all three variables drop significantly by removing the potential outliers, we can label them as outliers, however, before removing them we can check the outliers' data.

Checking the outliers' data.

```
Growth_of_urban_areas_rawdata_woutliers <- Growth_of_urban_areas_rawdata[all_outliers_Urbangrowth, ]

# Printing the data
Growth_of_urban_areas_rawdata_woutliers
```

Clearly, all the outliers are either on the higher end or the lower end through out the timeframe in the dataset, i.e. 2000-2020. Therefore, these records actually are not the outliers, and should not be removed from the data.

Mahalanobis distance method to deal with outliers in the 'Pollution' data

```
# Standardise the data
pollution_imputed_data_scaled <- scale(pollution_imputed_data[, c("pm10_concentration",
"pm25_concentration", "no2_concentration")])

# Compute Mahalanobis distance
Maha <- mahalanobis(pollution_imputed_data_scaled, colMeans(pollution_imputed_data_scaled),
cov(pollution_imputed_data_scaled))

# Calculating p-value for each Mahalanobis distance above
MahaPvalue <- pchisq(Maha, df=2, lower.tail = FALSE)
```

Printing no. of outliers

```
print(sum(MahaPvalue < 0.001))
```

Removing the outliers.

```
# Adding the p-values to the data
pollution_imputed_Maha <- cbind(pollution_imputed_data,MahaPvalue)

# Identifying the outliers
Mahaoutliers <- which(pollution_imputed_Maha$MahaPvalue < 0.001)

# Deleting the outliers
pollution_imputed_Maha <- pollution_imputed_Maha[-Mahaoutliers, ]

# Remove the p-values column and store back in the original dataset
pollution_imputed_data <- pollution_imputed_Maha[, !(names(pollution_imputed_Maha) == "MahaPvalue")]
```

```
# Re-standardising the data
pollution_imputed_data_scaled <- scale(pollution_imputed_data[, c("pm10_concentration",
"pm25_concentration", "no2_concentration")])
```

Clustering analysis to find air pollution hotspots

Estimating the number of clusters

```
# Plotting the Gap statistic
gap_stat_h <- clusGap(pollution_imputed_data_scaled, FUN = hcut, nstart = 25, K.max = 10, B = 50)
fviz_gap_stat(gap_stat_h)
```

Since, there is a jump in the above plot at k=3. The optimal number of clusters to be made is 3.

K-means clustering

```
# Perform clustering using k-means
set.seed(10)
clusters <- kmeans(pollution_imputed_data_scaled, 3)

# Add cluster results to the data
pollution_imputed_data$cluster <- clusters$cluster

# View the clustered data
head(pollution_imputed_data)
```

Summary of clusters

```
# Mean values of each cluster
clusters_averages <- aggregate(cbind(pm10_concentration, pm25_concentration, no2_concentration,
population) ~ cluster, data = pollution_imputed_data, FUN = "mean")

# Calculate the number of records in each cluster
cluster_counts <- pollution_imputed_data %>% group_by(cluster) %>% summarise(count = n())

# Merge the counts with the clusters_averages data frame
clusters_averages <- merge(clusters_averages, cluster_counts, by = "cluster")

# Print clusters' summary
print(clusters_averages)
```

Since, there are clear differences in terms of average PM10, PM2.5 and NO2 concentrations across three clusters. These clusters can be named as follows according to their level of concern.

1.) High Concern (Cluster 1): PM10, PM2.5 and NO2 levels are highest among all the clusters.
2.) 2.) Low Concern (Cluster 2): Concentration levels are lowest across clusters.
3.) 3.) Medium Concern (Cluster 3): PM10, PM2.5 and NO2 concentration levels are higher than that of cluster 2 but lower than cluster 1.

Adding these Concern levels as a new column in the data.

```
# Renaming the clusters
pollution_imputed_data <- pollution_imputed_data %>% mutate(concern_level =
ifelse(pollution_imputed_data$cluster==1, "High Concern", ifelse(pollution_imputed_data$cluster==2, "Low
Concern", "Medium Concern")))
```

Summarising the data by concern levels

```
pollution_imputed_data_summary <- pollution_imputed_data %>%
mutate(concern_level = factor(concern_level, levels = c("High Concern", "Medium Concern", "Low Concern")))
```

```
# Creating a summary dataset
pollution_imputed_data_summary <- pollution_imputed_data_summary %>% group_by(concern_level) %>%
summarise(pm10_concentration = mean(pm10_concentration), pm25_concentration =
mean(pm25_concentration), no2_concentration = mean(no2_concentration), population = mean(population))

# Printing the table
print(pollution_imputed_data_summary)
```

Dividing the data by years

```
Years <- sort(unique(pollution_imputed_data$year))
for (year in Years) {
assign(paste0("pollution_imputeddata_", year), pollution_imputed_data %>% filter(year == !!year))
}
```

Since, 2021 has only one record, we can skip this for the analysis.

Converting the data into xlsx

```
write.xlsx(pollution_imputed_data, "pollution_imputed_data.xlsx")
```

Note: Further graphs/ heatmaps are plotted using **Tableau**.

Exploratory Data Analysis Using the Socio-economic Data

(i) Year-on-Year Pollution Progression trends

Grouping pollution data by country, year and concern levels

```
# Average by country by year
pollution_imputed_data_by_country <- pollution_imputed_data %>% group_by(country_name, year) %>%
summarise(pm10_concentration = mean(pm10_concentration, na.rm = TRUE), pm25_concentration =
mean(pm25_concentration, na.rm = TRUE), no2_concentration = mean(no2_concentration, na.rm = TRUE))

# Average by year
pollution_imputed_data_by_year <- pollution_imputed_data %>% group_by(year) %>%
summarise(pm10_concentration = mean(pm10_concentration, na.rm = TRUE), pm25_concentration =
mean(pm25_concentration, na.rm = TRUE), no2_concentration = mean(no2_concentration, na.rm = TRUE))

# Average by concern level by year
pollution_imputed_data_by_concern <- pollution_imputed_data %>% group_by(concern_level, year) %>%
summarise(pm10_concentration = mean(pm10_concentration, na.rm = TRUE), pm25_concentration =
mean(pm25_concentration, na.rm = TRUE), no2_concentration = mean(no2_concentration, na.rm = TRUE),
avg_population = mean(population, na.rm = TRUE))
```

Overall pollution trends

```
# Plot Pollution trends
ggplot(pollution_imputed_data_by_year) +
geom_line(aes(x = year, y = pm10_concentration, color = "PM10")) +
geom_point(aes(x = year, y = pm10_concentration, color = "PM10")) +
geom_line(aes(x = year, y = pm25_concentration, color = "PM2.5")) +
geom_point(aes(x = year, y = pm25_concentration, color = "PM2.5")) +
geom_line(aes(x = year, y = no2_concentration, color = "NO2")) +
geom_point(aes(x = year, y = no2_concentration, color = "NO2")) +
scale_x_continuous(breaks = seq(2010, 2021, by = 1)) +
```

```
ylim(0, 150) +
labs(title = "Year-on-Year Progression of Pollutant Concentrations",
x = "Year",
y = "Concentration",
color = "Pollutant") +
theme_minimal() +
theme(panel.grid.major.x = element_line(color = "grey80"),
panel.grid.minor.x = element_blank(),
plot.title = element_text(hjust = 0.5))
```

Trends by country

```
# Plot for pm10_concentration
ggplot(pollution_imputed_data_by_country, aes(x = year, y = pm10_concentration, color = country_name)) +
geom_line() +
geom_point() +
scale_x_continuous(breaks = seq(2010, 2021, by = 1)) +
ylim(0, 210) +
labs(title = "Year-on-Year Progression of PM10 Concentration",
x = "Year",
y = "PM10 Concentration",
color = "Countries") +
theme_minimal() +
theme(panel.grid.major.x = element_line(color = "grey80"),
panel.grid.minor.x = element_blank(),
plot.title = element_text(hjust = 0.5))


# Plot for pm25_concentration
ggplot(pollution_imputed_data_by_country, aes(x = year, y = pm25_concentration, color = country_name)) +
geom_line() +
geom_point() +
scale_x_continuous(breaks = seq(2010, 2021, by = 1)) +
ylim(0, 210) +
labs(title = "Year-on-Year Progression of PM2.5 Concentration",
x = "Year",
y = "PM2.5 Concentration",
color = "Countries") +
theme_minimal() +
theme(panel.grid.major.x = element_line(color = "grey80"),
panel.grid.minor.x = element_blank(),
plot.title = element_text(hjust = 0.5))

# Plot for no2_concentration
ggplot(pollution_imputed_data_by_country, aes(x = year, y = no2_concentration, color = country_name)) +
geom_line() +
geom_point() +
scale_x_continuous(breaks = seq(2010, 2021, by = 1)) +
ylim(0, 210) +
labs(title = "Year-on-Year Progression of NO2 Concentration",
x = "Year",
y = "NO2 Concentration",
color = "Countries") +
theme_minimal() +
theme(panel.grid.major.x = element_line(color = "grey80"),
panel.grid.minor.x = element_blank(),
plot.title = element_text(hjust = 0.5))
```

Trends by Concern levels

```
# Plot for High Concern level
pollution_imputed_data_by_concern %>% filter(concern_level == "High Concern") %>% ggplot(aes(x = year))
geom_line(aes(y = pm10_concentration, color = "PM10 Concentration")) +
geom_line(aes(y = pm25_concentration, color = "PM2.5 Concentration")) +
geom_line(aes(y = no2_concentration, color = "NO2 Concentration")) +
geom_line(aes(y = avg_population / 100000, color = "Average Population (x00000)"), linetype = "dashed") +
scale_x_continuous(breaks = seq(2010, 2021, by = 1)) +
ylim(0, 250) +
labs(title = "Year-on-Year Pollution Progression (High Concern/ Hotspots)",
x = "Year",
y = "Concentration/ Population",
color = "Legend") +
theme_minimal() +
theme(panel.grid.major.x = element_line(color = "grey80"),
panel.grid.minor.x = element_blank(),
plot.title = element_text(hjust = 0.5)) +
scale_color_manual(values = c("PM10 Concentration" = "blue", "PM2.5 Concentration" = "red", "NO2
Concentration"= "green", "Average Population (x00000)" = "black"))

# Plot for Medium Concern level
pollution_imputed_data_by_concern %>% filter(concern_level == "Medium Concern") %>% ggplot(aes(x =
year)) geom_line(aes(y = pm10_concentration, color = "PM10 Concentration")) +
geom_line(aes(y = pm25_concentration, color = "PM2.5 Concentration")) +
geom_line(aes(y = no2_concentration, color = "NO2 Concentration")) +
geom_line(aes(y = avg_population / 100000, color = "Average Population (x00000)"), linetype = "dashed") +
scale_x_continuous(breaks = seq(2010, 2021, by = 1)) +
ylim(0, 250) +
labs(title = "Year-on-Year Pollution Progression (Medium Concern)",
x = "Year",
y = "Concentration/ Population",
color = "Legend") +
theme_minimal() +
theme(panel.grid.major.x = element_line(color = "grey80"),
panel.grid.minor.x = element_blank(),
plot.title = element_text(hjust = 0.5)) +
scale_color_manual(values = c("PM10 Concentration" = "blue", "PM2.5 Concentration" = "red", "NO2
Concentration"= "green", "Average Population (x00000)" = "black"))

# Plot for Low Concern level
pollution_imputed_data_by_concern %>% filter(concern_level == "Low Concern") %>% ggplot(aes(x = year)) +
geom_line(aes(y = pm10_concentration, color = "PM10 Concentration")) +
geom_line(aes(y = pm25_concentration, color = "PM2.5 Concentration")) +
geom_line(aes(y = no2_concentration, color = "NO2 Concentration")) +
geom_line(aes(y = avg_population / 100000, color = "Average Population (x00000)"), linetype = "dashed")
scale_x_continuous(breaks = seq(2010, 2021, by = 1)) +
ylim(0, 250) +
labs(title = "Year-on-Year Pollution Progression (Low Concern)",
x = "Year",
y = "Concentration/ Population",
color = "Legend") +
theme_minimal() +
theme(panel.grid.major.x = element_line(color = "grey80"),
panel.grid.minor.x = element_blank(),
plot.title = element_text(hjust = 0.5)) +
scale_color_manual(values = c("PM10 Concentration" = "blue", "PM2.5 Concentration" = "red", "NO2
Concentration"= "green", "Average Population (x00000)" = "black"))
```

(ii) Relation of Pollution and Socio-economic Data

One of the objectives of this study is to understand the relation between pollution levels and the socioeconomic situation. For this purpose it is essential to have them in one dataset. For better coverage, we will merge these socio-economic datasets with the 2020 pollution data since the socio-economic data is also from 2020. Growth of urban areas data also has 2010 data which can be merged with 2010 pollution data.

Applying operations to combine data for further analysis

```
# Preparing 2020 and 2010 pollution data to be merged with socio-economic datasets
pollution_data2020_tomerge <- pollution_imputeddata_2020 %>% rename('Country or Territory Name' = 'country_name', 'City Name' = 'city')
pollution_data2010_tomerge <- pollution_imputeddata_2010 %>% rename('Country or Territory Name' = 'country_name', 'City Name' = 'city')

# Merging 'Open & Green Spaces' data with 2020 pollution data
Open_and_green_spaces_merged <-
inner_join(Open_and_green_spaces_rawdata,pollution_data2020_tomerge[, c("Country or Territory Name", "City Name", "pm10_concentration", "pm25_concentration", "no2_concentration")], by = c("Country or Territory Name", "City Name"))

# Merging 'Urban transport' data with 2020 pollution data
Urban_transport_merged <- inner_join(Urban_transport_rawdata,pollution_data2020_tomerge[, c("Country or Territory Name", "City Name", "pm10_concentration", "pm25_concentration", "no2_concentration")], by = c("Country or Territory Name", "City Name"))

# Merging 'Growth of urban areas' data with 2010 and 2020 pollution data

# Merging 2020 pollution data and 2020 Growth data
Growth_of_urban_areas2020_merged <-
inner_join(Growth_of_urban_areas_rawdata,pollution_data2020_tomerge[, c("Country or Territory Name", "City Name", "pm10_concentration", "pm25_concentration", "no2_concentration")], by = c("Country or Territory Name", "City Name"))
# Selecting only the fields that are relevant for 2020
Growth_of_urban_areas2020_merged <- Growth_of_urban_areas2020_merged %>% rename('Data Year' = 'Data Year 3', 'Land consumption rate (LCR:%)' = 'Land consumption rate (LCR) Year 2 to Year 3 (%)', 'Population Growth Rate (PGR:%)' = 'Population Growth Rate (PGR) Year 2 to Year 3 (%)', 'Ratio of LCR to PGR (LCRPGR:Ratio)' = 'Ratio of Land Consumption Rate to Population Growth Rate (LCRPGR) Year 2 to Year 3 (Ratio)', 'Built Up area (m2 per person)' = 'Built Up area Per Capita Year 3 (m2 per person)') %>%
select(`Country or Territory Name`, `City Name`, `Data Year`, `Land consumption rate (LCR:%)`, `Population Growth Rate (PGR:%)`, `Ratio of LCR to PGR (LCRPGR:Ratio)`, `Built Up area (m2 per person)`, pm10_concentration, pm25_concentration, no2_concentration)

# Merging 2010 pollution data and 2010 Growth data
Growth_of_urban_areas2010_merged <-
inner_join(Growth_of_urban_areas_rawdata,pollution_data2010_tomerge[, c("Country or Territory Name", "City Name", "pm10_concentration", "pm25_concentration", "no2_concentration")], by = c("Country or Territory Name", "City Name"))
# Selecting only the fields that are relevant for 2010
Growth_of_urban_areas2010_merged <- Growth_of_urban_areas2010_merged %>% rename('Data Year' = 'Data Year 2', 'Land consumption rate (LCR:%)' = 'Land consumption rate (LCR) Year 1 to Year 2 (%)', 'Population Growth Rate (PGR:%)' = 'Population Growth Rate (PGR) Year 1 to Year 2 (%)', 'Ratio of LCR to PGR (LCRPGR:Ratio)' = 'Ratio of Land Consumption Rate to Population Growth Rate (LCRPGR) Year 1 to Year 2 (Ratio)', 'Built Up area (m2 per person)' = 'Built Up area Per Capita Year 2 (m2 per person)') %>%
select(`Country or Territory Name`, `City Name`, `Data Year`, `Land consumption rate (LCR:%)`, `Population Growth Rate (PGR:%)`, `Ratio of LCR to PGR (LCRPGR:Ratio)`, `Built Up area (m2 per person)`, pm10_concentration, pm25_concentration, no2_concentration)
```

```
# Combining 2020 and 2010 merged datasets
Growth_of_urban_areas_merged <- bind_rows(Growth_of_urban_areas2020_merged,
Growth_of_urban_areas2010_merged)

# Checking structure of these datasets
str(Open_and_green_spaces_merged)
str(Urban_transport_merged)
str(Growth_of_urban_areas_merged)
```

Relation of pollution with datapoints in 'Growth of urban areas' data

```
# Correlation matrix
GUA_Cor <- rcorr(as.matrix(select_if(Growth_of_urban_areas_merged, is.numeric)))

# Plotting the correlation matrix
corrplot(GUA_Cor$r, method = "color", type = "upper", tl.col = "black", tl.srt = 45, col =
colorRampPalette(c("red", "yellow", "green"))(100), main = "Growth of urban areas - Correlation Matrix", mar
= c(1, 1, 1, 1), tl.cex = 0.8)
```

```
# Printing the Correlation
print(GUA_Cor)
```

Now, we will be using mutiple regression models to understand these relations in detail. For multiple regression we will not use the 'Ratio of LCR to PGR (LCRPGR:Ratio)' variable since it is dependent on LCR and PGR.

Regression analysis on 'Growth of urban areas' data

```
# Multiple regression for PM10
GUA_pm10_model <- lm(pm10_concentration ~ `Land consumption rate (LCR:%)` + `Population Growth Rate
(PGR:%)` + `Built Up area (m2 per person)`, data = Growth_of_urban_areas_merged)

# Printing the model
summary(GUA_pm10_model)
confint(GUA_pm10_model)
```

```
# Add predicted values to the data frame
Growth_of_urban_areas_predict <- Growth_of_urban_areas_merged
Growth_of_urban_areas_predict$predicted_pm10 <- predict(GUA_pm10_model, newdata =
Growth_of_urban_areas_predict)

melted_data_GUA_PM10 <- melt(Growth_of_urban_areas_predict, id.vars = c("predicted_pm10"),
measure.vars = c("Land consumption rate (LCR:%)", "Population Growth Rate (PGR:%)", "Built Up area (m2 per
person)"))

# Extract the slopes
slopes_GUA_PM10 <- coef(GUA_pm10_model)[2:4]
slope_labels_GUA_PM10 <- c(paste("Slope:", round(slopes_GUA_PM10[1], 2)), paste("Slope:",
round(slopes_GUA_ PM10[2], 2)), paste("Slope:", round(slopes_GUA_ PM10[3], 2)))

# Create the faceted plot
ggplot(melted_data_GUA_PM10, aes(x = value, y = predicted_pm10, color = variable)) +
geom_smooth(method = "lm", se = FALSE) + facet_wrap(~ variable, scales = "free_x") + labs(title = "Predicted
PM10 Concentration vs Predictor Variables", x = "Predictor Variables", y = "Predicted PM10 Concentration",
color = "Predictor Variables:") + theme(legend.position = "bottom") + scale_color_manual(values = c("Land
consumption rate (LCR:%)" = "blue", "Population Growth Rate (PGR:%)" = "red", "Built Up area (m2 per
person)" = "purple"), labels = c("Land consumption rate (LCR:%)" = "LCR (%)", "Population Growth Rate
(PGR:%)" = "PGR (%)", "Built Up area (m2 per person)" = "Built Up Area (m2/person)")) + geom_text(data =
data.frame(variable = c("Land consumption rate (LCR:%)", "Population Growth Rate (PGR:%)", "Built Up area
```

```
(m2 per person)"), label = slope_labels_GUA_PM10), aes(x = Inf, y = Inf, label = label), hjust = 1.1, vjust = 2, size = 4, color = "black")
```

```
# Multiple regression for PM2.5
GUA_pm2.5_model <- lm(pm25_concentration ~ `Land consumption rate (LCR:%)` + `Population Growth Rate (PGR:%)` + `Built Up area (m2 per person)`, data = Growth_of_urban_areas_merged)

# Printing the model
summary(GUA_pm2.5_model)
confint(GUA_pm2.5_model)
```

```
# Add predicted values to the data frame
Growth_of_urban_areas_predict$predicted_pm2.5 <- predict(GUA_pm2.5_model, newdata = Growth_of_urban_areas_predict

melted_data_GUA_PM2.5 <- melt(Growth_of_urban_areas_predict, id.vars = c("predicted_pm2.5"), measure.vars = c("Land consumption rate (LCR:%)", "Population Growth Rate (PGR:%)", "Built Up area (m2 per person)"))

# Extract the slopes
slopes_GUA_PM25 <- coef(GUA_pm2.5_model)[2:4]
slope_labels_GUA_PM25 <- c(paste("Slope:", round(slopes_GUA_PM25[1], 2)), paste("Slope:", round(slopes_GUA_ PM25[2], 2)), paste("Slope:", round(slopes_GUA_ PM25[3], 2)))

# Create the faceted plot
ggplot(melted_data_GUA_PM2.5, aes(x = value, y = predicted_pm2.5, color = variable)) + geom_smooth(method = "lm", se = FALSE) + facet_wrap(~ variable, scales = "free_x") + labs(title = "Predicted PM2.5 Concentration vs Predictor Variables", x = "Predictor Variables", y = "Predicted PM2.5 Concentration", color = "Predictor Variables:") + theme(legend.position = "bottom") + scale_color_manual(values = c("Land consumption rate (LCR:%)" = "blue", "Population Growth Rate (PGR:%)" = "red", "Built Up area (m2 per person)" = "purple"), labels = c("Land consumption rate (LCR:%)" = "LCR (%)", "Population Growth Rate (PGR:%)" = "PGR (%)", "Built Up area (m2 per person)" = "Built Up Area (m2/person)")) + geom_text(data = data.frame(variable = c("Land consumption rate (LCR:%)", "Population Growth Rate (PGR:%)", "Built Up area (m2 per person)"), label = slope_labels_GUA_PM25), aes(x = Inf, y = Inf, label = label), hjust = 1.1, vjust = 2, size = 4, color = "black")
```

```
# Multiple regression for NO2
GUA_no2_model <- lm(no2_concentration ~ `Land consumption rate (LCR:%)` + `Population Growth Rate (PGR:%)` + `Built Up area (m2 per person)`, data = Growth_of_urban_areas_merged)

# Printing the model
summary(GUA_no2_model)
confint(GUA_no2_model)
```

```
# Add predicted values to the data frame
Growth_of_urban_areas_predict$predicted_no2 <- predict(GUA_no2_model, newdata = Growth_of_urban_areas_predict)

melted_data_GUA_NO2 <- melt(Growth_of_urban_areas_predict, id.vars = c("predicted_no2"), measure.vars = c("Land consumption rate (LCR:%)", "Population Growth Rate (PGR:%)", "Built Up area (m2 per person)"))

# Extract the slopes
slopes_GUA_NO2 <- coef(GUA_no2_model)[2:4]
slope_labels_GUA_NO2 <- c(paste("Slope:", round(slopes_GUA_NO2[1], 2)), paste("Slope:", round(slopes_GUA_ NO2[2], 2)), paste("Slope:", round(slopes_GUA_ NO2[3], 2)))
```

```
# Create the faceted plot
ggplot(melted_data_GUA_NO2, aes(x = value, y = predicted_no2, color = variable)) + geom_smooth(method = "lm", se = FALSE) + facet_wrap(~ variable, scales = "free_x") + labs(title = "Predicted NO2 Concentration vs Predictor Variables", x = "Predictor Variables", y = "Predicted NO2 Concentration", color = "Predictor Variables:") + theme(legend.position = "bottom") + scale_color_manual(values = c("Land consumption rate (LCR:%)" = "blue", "Population Growth Rate (PGR:%)" = "red", "Built Up area (m2 per person)" = "purple"), labels = c("Land consumption rate (LCR:%)" = "LCR (%)", "Population Growth Rate (PGR:%)" = "PGR (%)", "Built Up area (m2 per person)" = "Built Up Area (m2/person)")) + geom_text(data = data.frame(variable = c("Land consumption rate (LCR:%)", "Population Growth Rate (PGR:%)", "Built Up area (m2 per person)"), label = slope_labels_GUA_NO2), aes(x = Inf, y = Inf, label = label), hjust = 1.1, vjust = 2, size = 4, color = "black")
```

Relation of pollution with datapoints in 'Urban transport' data

```
# Renaming 'Share of urban population with convenient access to public transport (%)'
Urban_transport_merged <- Urban_transport_merged %>% rename(`Access to Public Transport` = `Share of urban population with convenient access to public transport (%)`)

# Correlation matrix
UT_Cor <- rcorr(as.matrix(Urban_transport_merged %>% select_if(is.numeric) %>% select(-`Data Reference Year`)))

UT_Cor$r[is.infinite(UT_Cor$r)] <- NA

# Plotting the correlation matrix
corrplot(UT_Cor$r, method = "color", type = "upper", tl.col = "black", tl.srt = 45, col = colorRampPalette(c("red", "yellow", "green"))(100), main = "Urban transport - Correlation Matrix", mar = c(1, 1, 1, 1), tl.cex = 0.8)
```

```
# Printing the Correlation
print(UT_Cor)
```

Regression analysis on 'Urban transport' data

```
# Regression analysis for PM10
UT_pm10_model <- lm(pm10_concentration ~ `Access to Public Transport`, data = Urban_transport_merged)

# Printing the model
summary(UT_pm10_model)
confint(UT_pm10_model)
```

```
# Add predicted values to the data frame
Urban_transport_predict <- Urban_transport_merged
Urban_transport_predict$predicted_pm10 <- predict(UT_pm10_model, newdata = Urban_transport_predict)

melted_data_UT_PM10 <- melt(Urban_transport_predict, id.vars = c("predicted_pm10"), measure.vars = "Access to Public Transport")

# Extract the slope
slope_UT_PM10 <- coef(UT_pm10_model)[2]

# Create the faceted plot
ggplot(melted_data_UT_PM10, aes(x = value, y = predicted_pm10)) +
geom_smooth(method = "lm", se = FALSE) +
labs(title = "Predicted PM10 Levels vs Urban Population with Access to Public Transport",
x = "% of Urban Population with Access to Public Transport",
y = "Predicted PM10 Concentration") +
theme(legend.position = "none") +
annotate("text", x = Inf, y = Inf, label = paste("Slope:", round(slope_UT_PM10, 2)), hjust = 1.1, vjust = 2, size = 4, color = "black")
```

```r
# Regression analysis for PM2.5
UT_pm2.5_model <- lm(pm25_concentration ~ `Access to Public Transport`, data = Urban_transport_merged)

# Printing the model
summary(UT_pm2.5_model)
confint(UT_pm2.5_model)
```

```r
# Add predicted values to the data frame
Urban_transport_predict$predicted_pm2.5 <- predict(UT_pm2.5_model, newdata = Urban_transport_predict)

melted_data_UT_PM2.5 <- melt(Urban_transport_predict, id.vars = c("predicted_pm2.5"), measure.vars = "Access to Public Transport")

# Extract the slope
slope_UT_PM2.5 <- coef(UT_pm2.5_model)[2]

# Create the faceted plot
ggplot(melted_data_UT_PM2.5, aes(x = value, y = predicted_pm2.5)) +
geom_smooth(method = "lm", se = FALSE) +
labs(title = "Predicted PM2.5 Levels vs Urban Population with Access to Public Transport",
x = "% of Urban Population with Access to Public Transport",
y = "Predicted PM2.5 Concentration") +
theme(legend.position = "none") +
annotate("text", x = Inf, y = Inf, label = paste("Slope:", round(slope_UT_PM2.5, 2)), hjust = 1.1, vjust = 2, size = 4, color = "black")
```

```r
# Regression analysis for NO2
UT_no2_model <- lm(no2_concentration ~ `Access to Public Transport`, data = Urban_transport_merged)

# Printing the model
summary(UT_no2_model)
confint(UT_no2_model)
```

```r
# Add predicted values to the data frame
Urban_transport_predict$predicted_no2 <- predict(UT_no2_model, newdata = Urban_transport_predict)

melted_data_UT_NO2 <- melt(Urban_transport_predict, id.vars = c("predicted_no2"), measure.vars = "Access to Public Transport")

# Extract the slope
slope_UT_NO2 <- coef(UT_no2_model)[2]

# Create the faceted plot
ggplot(melted_data_UT_NO2, aes(x = value, y = predicted_no2)) +
geom_smooth(method = "lm", se = FALSE) +
labs(title = "Predicted NO2 Levels vs Urban Population with Access to Public Transport",
x = "% of Urban Population with Access to Public Transport",
y = "Predicted NO2 Concentration") +
theme(legend.position = "none") +
annotate("text", x = Inf, y = Inf, label = paste("Slope:", round(slope_UT_NO2, 2)), hjust = 1.1, vjust = 2, size = 4, color = "black")
```

Relation of pollution with datapoints in 'Open & Green Spaces' data

```
# Renaming the variables for better visualisation
Open_and_green_spaces_merged <- Open_and_green_spaces_merged %>% rename(`Open Space Share` = `Average share of the built-up area of cities that is open space for public use for all (%) [a]`,`Access to Open Spaces` = `Average share of urban population with convenient access to open public spaces (%) [b]`)

# Correlation matrix
OGS_Cor <- rcorr(as.matrix(Open_and_green_spaces_merged %>% select_if(is.numeric) %>% select(-`Data Reference Year`)))

OGS_Cor$r[is.infinite(OGS_Cor$r)] <- NA

# Plotting the correlation matrix
corrplot(OGS_Cor$r, method = "color", type = "upper", tl.col = "black", tl.srt = 45, col = colorRampPalette(c("red", "yellow", "green"))(100), main = "Open & Green Spaces - Correlation Matrix", mar = c(1, 1, 1, 1), tl.cex = 0.8)
```

```
# Printing the Correlation
print(OGS_Cor)
```

Regression analysis on 'Open & Green Spaces' data

```
# Regression analysis for PM10
OGS_pm10_model <- lm(pm10_concentration ~ `Open Space Share` + `Access to Open Spaces`, data = Open_and_green_spaces_merged)

# Printing the model
summary(OGS_pm10_model)
confint(OGS_pm10_model)
```

```
# Add predicted values to the data frame
Open_and_green_spaces_predict <- Open_and_green_spaces_merged
Open_and_green_spaces_predict$predicted_pm10 <- predict(OGS_pm10_model, newdata = Open_and_green_spaces_predict

melted_data_OGS_PM10 <- melt(Open_and_green_spaces_predict, id.vars = c("predicted_pm10"), measure.vars = c("Open Space Share", "Access to Open Spaces"))

# Extract the slopes
slopes_OGS_PM10 <- coef(OGS_pm10_model)[2:3]
slope_labels_OGS_PM10 <- c(paste("Slope:", round(slopes_OGS_PM10[1], 2)), paste("Slope:", round(slopes_OGS_ PM10[2], 2)))

# Create the faceted plot
ggplot(melted_data_OGS_PM10, aes(x = value, y = predicted_pm10, color = variable)) + geom_smooth(method = "lm", se = FALSE) + facet_wrap(~ variable, scales = "free_x") + labs(title = "Predicted PM10 Levels vs Predictor Variables", x = "Predictor Variables", y = "Predicted PM10 Concentration", color = "Predictor Variables:") + theme(legend.position = "bottom") + scale_color_manual(values = c("Open Space Share" = "blue", "Access to Open Spaces" = "red"), labels = c("Open Space Share" = "Open Space Share", "Access to Open Spaces" = "Access to Open Spaces")) + geom_text(data = data.frame(variable = c("Open Space Share", "Access to Open Spaces"), label = slope_labels_OGS_PM10), aes(x = Inf, y = Inf, label = label), hjust = 1.1, vjust = 2, size = 4, color = "black")
```

```
# Regression analysis for PM2.5
OGS_pm2.5_model <- lm(pm25_concentration ~ `Open Space Share` + `Access to Open Spaces`, data =
Open_and_green_spaces_merged)

# Printing the model
summary(OGS_pm2.5_model)
confint(OGS_pm2.5_model)
```

```
# Add predicted values to the data frame
Open_and_green_spaces_predict <- Open_and_green_spaces_merged
Open_and_green_spaces_predict$predicted_pm2.5 <- predict(OGS_pm2.5_model, newdata =
Open_and_green_spaces_ predict)

melted_data_OGS_PM2.5 <- melt(Open_and_green_spaces_predict, id.vars = c("predicted_pm2.5"),
measure.vars = c("Open Space Share", "Access to Open Spaces"))

# Extract the slopes
slopes_OGS_PM2.5 <- coef(OGS_pm2.5_model)[2:3]
slope_OGSPM2.5_labels <- c(paste("Slope:", round(slopes_OGS_PM2.5[1], 2)), paste("Slope:", round(slopes_
OGS_PM2.5[2], 2)))

# Create the faceted plot
ggplot(melted_data_OGS_PM2.5, aes(x = value, y = predicted_pm2.5, color = variable)) +
geom_smooth(method = "lm", se = FALSE) + facet_wrap(~ variable, scales = "free_x") + labs(title = "Predicted
PM2.5 Levels vs Predictor Variables", x = "Predictor Variables", y = "Predicted PM2.5 Concentration", color =
"Predictor Variables:") + theme(legend.position = "bottom") + scale_color_manual(values = c("Open Space
Share" = "blue", "Access to Open Spaces" = "red"), labels = c("Open Space Share" = "Open Space Share",
"Access to Open Spaces" = "Access to Open Spaces")) + geom_text(data = data.frame(variable = c("Open
Space Share", "Access to Open Spaces"), label = slope_OGSPM2.5_labels), aes(x = Inf, y = Inf, label = label),
hjust = 1.1, vjust = 2, size = 4, color = "black")
```

```
# Regression analysis for NO2
OGS_no2_model <- lm(no2_concentration ~ `Open Space Share` + `Access to Open Spaces`, data =
Open_and_green_spaces_merged)

# Printing the model
summary(OGS_no2_model)
confint(OGS_no2_model)
```

```
# Add predicted values to the data frame
Open_and_green_spaces_predict <- Open_and_green_spaces_merged
Open_and_green_spaces_predict$predicted_no2 <- predict(OGS_no2_model, newdata =
Open_and_green_spaces_predict)

melted_data_OGS_NO2 <- melt(Open_and_green_spaces_predict, id.vars = c("predicted_no2"), measure.vars
= c("Open Space Share", "Access to Open Spaces"))

# Extract the slopes
slopes_OGS_NO2 <- coef(OGS_no2_model)[2:3]
slope_OGSNO2_labels <- c(paste("Slope:", round(slopes_OGS_NO2[1], 2)), paste("Slope:", round(slopes_OGS_
NO2[1], 2)))
```

```
# Create the faceted plot
ggplot(melted_data_OGS_NO2, aes(x = value, y = predicted_no2, color = variable)) + geom_smooth(method = "lm", se = FALSE) + facet_wrap(~ variable, scales = "free_x") + labs(title = "Predicted NO2 Levels vs Predictor Variables", x = "Predictor Variables", y = "Predicted NO2 Concentration", color = "Predictor Variables:") + theme(legend.position = "bottom") + scale_color_manual(values = c("Open Space Share" = "blue", "Access to Open Spaces" = "red"), labels = c("Open Space Share" = "Open Space Share", "Access to Open Spaces" = "Access to Open Spaces")) + geom_text(data = data.frame(variable = c("Open Space Share", "Access to Open Spaces"), label = slope_OGSNO2_labels), aes(x = Inf, y = Inf, label = label), hjust = 1.1, vjust = 2, size = 4, color = "black")
```

Relation of pollution with population

```
# Correlation matrix
Pol_Cor <- rcorr(as.matrix(select_if(pollution_imputed_data, is.numeric)))

# Plotting the correlation matrix
corrplot(Pol_Cor$r, method = "color", type = "upper", tl.col = "black", tl.srt = 45, col = colorRampPalette(c("red", "yellow", "green"))(100), main = "Pollution data - Correlation Matrix", mar = c(1, 1, 1, 1), tl.cex = 0.8)
```

```
# Printing the Correlation
print(Pol_Cor)
```

Regression analysis on 'Pollution' data

```
# Regression analysis for PM10
Poll_pm10_model <- lm(pm10_concentration ~ population, data = pollution_imputed_data)

# Printing the model
summary(Poll_pm10_model)
confint(Poll_pm10_model)
```

```
# Regression analysis for PM2.5
Poll_pm2.5_model <- lm(pm25_concentration ~ population, data = pollution_imputed_data)

# Printing the model
summary(Poll_pm2.5_model)
confint(Poll_pm2.5_model)
```

```
# Regression analysis for NO2
Poll_no2_model <- lm(no2_concentration ~ population, data = pollution_imputed_data)

# Printing the model
summary(Poll_no2_model)
confint(Poll_no2_model)
```