University of Westminster

Team 3

Water - contaminants and levels

2024

Contents

Project Plan....3
Project charter....3
Project Background....6
Project Description....6

Project Importance....6
Problem Structuring Method....7
Project Scope....10
Timeline....15
Communication plan....15
Background research....20
Data set analysis....21
Data evaluation....21
Data pre-processing....26
Raw data overview....26
Data relevance and selection....30
Applied methodology in data pre-processing....31
Statistical modelling and error margins....35
Analysis and recommendations....38
Wastewater discharges impact on Health....38
Wastewater discharges impact on Environment....41
Random Forest Regression....44
Wastewater impact on society....46
Wastewater impact on carbon density....48
Insights and recommendations....49
Individual reflections....52
References....69

Project Plan

Project charter

Company: Koru Impact Solutions

Client Details :

Name	Email	Phone	Preferred Contact Method
Kate Barnard	kate@koruimpactsolutions.com	07961847528	email

Module Leader Details:

Name	Email	Phone	Preferred Contact Method
Salma Chahed	S.Chahed@westminster.ac.uk		email

Stakeholders:

Koru Impact Solutions, customers (B2B fund managers), regulatory bodies (FCA, SRI), competitors, potential investors, public.

Team Details:

Name	Email	Phone	Role/ Responsibilities
Laurita Kunickaite	w1947567@my.westminster.ac.uk	07384707050	Project manager: Create and keep a clear project plan of action

			Organizing and minute team meetings, providing the plan for the meetings Provide updates to the client on the team’s progress Provide insights and recommendations in a report for business executives
Om Sadigale	w1943544@my.westminster.ac.uk	+919076249855	<u>Data auditor:</u> Identify inherent risks or biases of data set affecting its accuracy Ensure reliability of data source by means such as interrogating data collection process
Yahya Habib	w1948192@my.westminster.ac.uk	+447497062433	<u>Developer of statistical model:</u> Creation of statistical model to measure the margins of error against key metrics Assist the data auditor in Improving quality of sourced data
Anas Kagigi	w1914597@my.westminster.ac.uk	+447576801893	<u>Data researcher:</u> Research additional data resources to contribute and

			create more wholistic data set to improve data accuracy
Faisal Qaderi	w1897498@my.westminster.ac.uk	+447383337940	<u>Python programmer:</u> Cleaning and transforming data for analysis using Python Visualise the data that have been cleansed for further analysis Advised on decision-making process, using Python data mentioned above
Ayaan Iqbal	w1948385@my.westminster.ac.uk	+447709804137	<u>Implementer:</u> Inputs the analysed data and reviews against the key lenses of health, society, environment and carbon intensity

Document Control:

Created on: 08 February 2024

Last updated on: 30 March 2024

Updated by: Laurita Kunickaite

Estimates:

Expected Start Date: 06 February 2024

Expected End Date: 28 March 2024

Project Background

Koru Impact Solutions is a revolutionizing organization with its AI-driven platform, which allows companies to partner with impactful projects worldwide that align with their strategic investments and sustainability goals. The company aims to replace the laborious process of producing impact reports with an automated digital solution, leveraging APIs and data sets, to generate enhanced reports for fund managers. Reports are being directed by regulatory bodies such as the Financial Conduct Authority (FCA) and Socially Responsible Investing (SRI) frameworks.

The challenge is ensuring the accuracy, reliability and consistency of the data used in creating those reports. Moreover, expanding the process of reporting meaningful insights into the environmental and sustainability impacts of investment portfolios.

Project Description

The project goal is to address the pollution issue of water contaminants and their levels and create an impact report for fund managers. Koru Impact Solutions addresses the significance of water quality as a key environmental metric affecting humans' health and ecosystem integrity.

Objectives of this project include integrating the provided data set, applying the created statistical model and providing insights and recommendations for the client. The main aim is to highlight the increasing importance of environmental, social and governance (ESG) considerations in investments. Also, a broader understanding of water contamination risks.

Overall, by bringing attention to water pollution and its causes in the report, the project aims to encourage fund managers to make informative decisions supporting sustainable water management, contributing to environmental stewardship goals.

Project Importance

Conducting this project is crucially important for Koru Impact Solutions as it meets the company's aim to automate impact reports and promote sustainable investing practices to its customers. By creating a digital auditing system focusing in this case – water contaminants, the organization expands its position as a leader and enables purpose-led

investment. This project helps the company to stand out from competitors by offering quicker and more innovative solutions that fulfil the needs of B2B clients and fund managers.

Moreover, Koru Impact Solutions demonstrates its determination to environmental and social responsibility. The successful implementation of this project not only strengthens the organization's market position but also promotes transparency, accountability and positive environmental outcomes in the financial sector. Ultimately, the company by prioritizing this project solidifies its reputation as a trusted partner and provides meaningful impact in the industry.

Problem Structuring Method

The core problem of Koru Impact:

Improve accuracy and consistency of impact reporting for sustainability metrics.

1. **Human activity systems:**
   1.1. Koru impact team:
   - Eager to automate audit reports and promote sustainable investing practices

   1.2. Fund managers:
   - Expects accurate and reliable impact reports which they can use for investment decision–making
   - Reliant on impact metrics to demonstrate sustainability

   1.3. Regulatory Bodies (FCA, SDR regulations):
   - Setting guidelines and regulations for impact reporting in the financial sector
   - Promotes transparency and accountability in sustainable investing

   1.4. Competitors:
   - Such as MSCI, Sustainalytics, and Impacted cubed, offer similar solutions for impact reporting,
   - May serve as benchmarks for Koru Impact Solutions, motivating them to improve their own offerings and stay ahead in the market

   1.5. Investors:

- Allocate capital to investment opportunities
- May have varying preferences and priorities regarding sustainability metrics, such as carbon emissions reduction, social impact, or biodiversity conservation
- Koru Impact Solutions must cater to investors' needs and preferences, providing them with transparent and actionable insights to support their sustainability goals and investment strategies.

2. Issues:

2.1. Lack of data accuracy and consistency:

- Inherent risks and biases in provided datasets
- Difficult integration of diverse global datasets
- Difficulty in concluding provided datasets

2.2. Challenges in statistical modelling:

- Limited experience in developing accurate statistical model for error margin assessment
- Necessity for a standardized approach to evaluate error margins across different environmental metrics

2.3. Integration of external data sets:

- Finding reliable public data sets to enhance accuracy
- Ensuring compatibility and consistency in data integration

3. Root definition (CATWOE):

3.1. Client:

- Fund managers defined by regulatory bodies
- Investors, stakeholders in financial sector

3.2. Actors:

- Koru Impact solution – responsible for developing the automated impact reporting tool
- Students – data evaluation, preprocessing and providing actionable insights and recommendations

3.3. Transformation:

- Objective: uncleaned data not suitable for impact reporting
- Process: evaluating data sources, pre-processing and integrating data, developing statistical models, integrating additional datasets, analysing outputs, and providing insights and recommendations.

- Outcome: cleaned data suitable for impact reporting that provides accurate and reliable insights for fund managers
- Regulatory bodies – set guidelines and regulations for impact reporting
- Stakeholders: fund managers, investors, regulatory bodies

3.4. World view:

- Importance of promoting sustainability and providing innovative solutions which enable purpose-led investment.

3.5. Owner:

- Koru Impact Solutions

3.6. Environment:

- Internal: organizational culture, resources and capabilities
- External: regulatory requirements, market trends, competition, available data sources, technological advancements and stakeholders' expectations.
- Ethical: such as MSCI, Sustainalytics, and Impacted cubed, offer similar solutions for impact reporting, may serve as benchmarks for Koru Impact Solutions, motivating them to improve their own offerings and stay ahead in the market

4. Desirable changes:

4.1. Improved Data:

- Implemented data quality measures and data pre-processing tools

4.2. Enhanced collaboration:

- Engaging with stakeholders for feedback and collaboration in data integration efforts

Project Scope

Project deliverables	Project Outcomes	Key activities to achieve the outcome	Activity lead
Comprehensive data evaluation	**Understanding of water contaminant issues and identification of biases in data**	1. Identify inherent risks or biases of the data and assessing quality of it 2. Conduct the reliability of the data source 3. Identify key trends and patterns of concern related to water contaminants 4. Relevant background research for water contaminants and impact	Om Sadigale, Faisal Qaderi
Pre-processed and integrated data sets	**1. Cleaning and integrating data for global analysis**	1.1. Cleaning and transforming data 1.2. Standardize data formats and units 1.3. Aggregate data to the global level	Faisal Qaderi, Yahya Habib, Om Sadigale
	2. Implement a statistical model for error margin measurement	2.1. Select the appropriate statistical model for measuring the error margins 2.2. Use algorithms and formulas to calculate error margins based on data precision, geographic variability, and other relevant factors. 2.3. Test and validate the statistical model 2.4. Improve and optimize the model based on testing results and improve accuracy and reliability.	Faisal Qaderi, Yahya Habib
Additional data set integration and aggregated outputs analysis	**1. Integrate diverse data sources for improvement of the accuracy of the metric**	1.1. Find relevant data sources/ literature 1.2. Assess the quality and relevance of new data sets 1.3. Aggregate new data with existing data for analysis 1.4. Analyse the impact of the integration on error margins	Anas Kagigi, Ayaan Iqbal
	2. Assess the impact of water contaminants	2.1. Apply results through health, society, environment, and carbon intensity lenses 2.2. Assess impact on each field	Ayaan Iqbal,

		2.3. Summarize key findings and trends	Anas Kagigi
Insights and recommendations report	**Provide key insights and recommendations based on analysis**	1. Summarize key findings from the analysis 2. Provide actionable insights and recommendations 3. Combining created reports and presentations	Laurita Kunickaite
Resources required		Provided data set and additional data sets, project plan, interview with a stakeholder, advanced analytical tools (Python), platform for storing code (Github), timeline planning software (Gantt), human resources (students in this project), sufficient time, communication tools (WhatsApp, email), collaboration with module leader	
Project Constraints		Tight project timelines, lack/availability of high-quality additional data sources, lack of experience in statistical models and project management.	

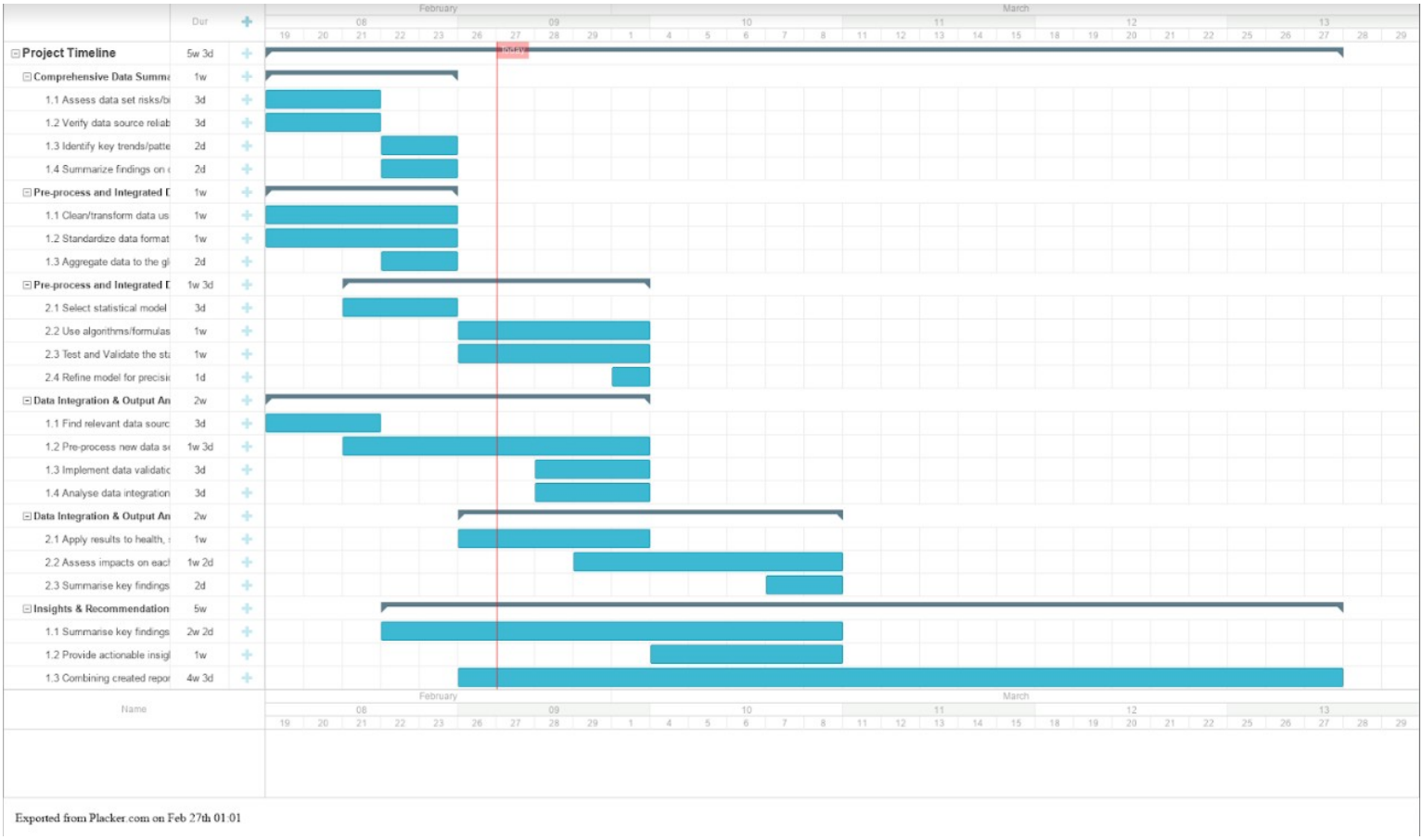
Risks		**Likelihood and impact score**	**Mitigation**
	Data quality and availability	**High**	1. **Proceed through data validation:** implement code validation processes and manual checks where necessary 2. **Implement data cleaning and processing techniques to ensure accuracy of data:** use algorithms and tools to clean, and preprocess data to ensure its accuracy. 3. **Collaborate with reliable data provider:** Assess data providers thoroughly, get

			feedback from the client
	Technical skills	**Medium**	1. **Ensure regular communication with team members and raise concerns:** Foster an open communication culture to address skill gaps and concerns promptly 2. **Seek additional information:** Encourage team members to use lecture notes, online resources to enhance skills.
	Experience	**Medium**	1. **Identify and address skill gaps:** conduct regular skill assessments and provide the support needed 2. **Prioritize tasks based on team members' skills:** Assign tasks according to individual strengths and provide support where needed
	Stakeholder management	**High**	1. **Establish clear channels of communication:** set up regular meetings, use collaboration tools, and define escalation path. 2. **Provide regular**

			updates: share progress reports, milestones achieved, and upcoming plans consistently. 3. **Address stakeholder concerns promptly:** prioritize and resolve issues swiftly to maintain stakeholder confidence. 4. **Ensure collaborative and transparent project environment:** foster trust by involving stakeholders in decision-making and being transparent about project challenges
	Timeline delays	**High**	1. **Develop a realistic project timeline:** account for potential delays and buffer time for unexpected issues. 2. **Regularly monitor project progress:** communicate with team about progress with tasks and milestones 3. **Identify and address issues early:** encourage proactive problem-solving and escalate issues promptly

			4. **Adjust timeline as needed:** Reassess and update the timeline regularly based on project progress
	Project scope	**High**	1. **Identify clear project objectives and boundaries:** document and communicate project scope clearly to all stakeholders 2. **Regularly review and prioritize project tasks:** conduct frequent reviews to ensure alignment with project objectives and adjust priorities accordingly. 3. **Gain stakeholders' approval for any changes:** implement a change management process to evaluate and approve scope changes systematically.

Timeline



Communication plan

Team Meetings:		
Week 1: Orientation- initial team meeting		
Date and location	**Meeting minutes**	**Team members attended**
06/02/24 21:00 -Zoom	Talked about project brief,	Laurita Kunickaite,

	and identified technical skills, personal strengths and weaknesses, which helped to assign the roles and responsibilities for each team member. Start putting together slides for team charter submission on 8/02/24	Om Sadigale, Faisal Qaderi, Yahya Habib, Ayaan Iqbal, Anas Kagigi
07/02/24 21:00 – Zoom	Reviewed draft slides for team charter. Discussed responsibilities for each team member role.	Laurita Kunickaite, Om Sadigale, Faisal Qaderi, Yahya Habib, Ayaan Iqbal, Anas Kagigi
Week 2: Project Briefing with Industry Client		
Date and location	**Meeting minutes**	**Team members attended**
11/02/24 20:00 - Zoom	Preparation for the meeting with the client on 13/02/24: discussed the questions which could be asked. Analysis of customer/ project brief. Start talking about the project plan, and discuss about information found on the Practera website. We looked into the provided data sets and started making activity plan for what needs to be done about data preprocessing and data validation	Laurita Kunickaite, Om Sadigale, Faisal Qaderi, Yahya Habib, Ayaan Iqbal, Anas Kagigi
13/02/24 21:00 – Zoom	Discussed the meeting with the client and ensured that we all understood what is the project about and what is our task.	Laurita Kunickaite, Om Sadigale, Faisal Qaderi, Yahya Habib, Ayaan Iqbal,

	Finalising and discussing the project plan for submission on 15/02/24. Clarifying all concerns about the task assigned to each member.	Anas Kagigi
Week 3: Data preparation, analysis & initial recommendations		
Date and location	**Meeting minutes**	**Team members attended**
18/02/24 21:00 - Zoom	Discussed data validation and data pre-processing results. Draw some insights. Team members suggested some improvements in pre-processing data further. Discussed additional data sources found and what insights we can implement in our analysed data. Brainstormed assumptions about processed data results	Laurita Kunickaite, Om Sadigale, Faisal Qaderi, Yahya Habib, Ayaan Iqbal, Anas Kagigi
20/02/24 21:00 – Zoom	Started drafting technical reports about data pre-processing and data validation. Summarising findings and additional relevant information from external data sources. Talked with the team about the draft report submission, and what needs to be included. Discussed the answer of questions asked from client.	Laurita Kunickaite, Om Sadigale, Faisal Qaderi, Yahya Habib, Ayaan Iqbal, Anas Kagigi

Week 4: Data analysis & initial recommendations		
Date and location	**Meeting minutes**	**Team members attended**
25/02/24, 21:00 - Zoom	Discussed data validation results and external data sources. Talked trough insights found from external data sources, which can be integrated with our data set. Discussed any queries regarding writing the insights and visualisation of data.	Laurita Kunickaite, Om Sadigale, Ayaan Iqbal, Anas Kagigi
27/02/24 21:00 – Zoom	Discussed a technical report and some insights into the data set. Talked through any raised issues within the team. Clarified further steps and checked if the team was following the project plan.	Laurita Kunickaite, Om Sadigale, Faisal Qaderi, Yahya Habib, Ayaan Iqbal, Anas Kagigi
Week 5: Bringing together project draft report		
Date and location	**Meeting minutes**	**Team members attended**
(03/03/24, 21:00 - Zoom	Discussed the feedback received from the module leader. Brainstormed the ideas of how the project could be improved. Analysed the issues with the data pre-processing part. Talk about possible insights	Laurita Kunickaite, Om Sadigale, Ayaan Iqbal, Anas Kagigi

	from analysed data and how it can be correlated with external data. Agreed to prepare questions for the meeting with the module leader. To clarify all uncertainties.	
05/03/24 21:00 - Zoom)	Meeting outcomes were discussed Agreed the new approach for data pre-processing part and deadline for this technical part	Laurita Kunickaite, Om Sadigale, Faisal Qaderi, Yahya Habib, Ayaan Iqbal, Anas Kagigi
Week 6: Draft report		
Date and location	**Meeting minutes**	**Team members attended**
(10/03/24, 21:00 – Zoom)	Discussion and analysis for data pre-processing final draft. Brainstormed about assumptions that could be made from the results we got. Discussed team progress and plan for the next week before a draft submission on 18th March	Laurita Kunickaite, Om Sadigale, Faisal Qaderi, Yahya Habib, Ayaan Iqbal,
(12/03/24 13:30 – Module seminar)	Talked through was what achieved with data pre-processing and look into report draft Agreed that more visualisation needed and how they are going to me made Clarified the task for the rest of week until the next meeting	Laurita Kunickaite, Om Sadigale, Faisal Qaderi, Yahya Habib, Ayaan Iqbal, Anas Kagigi

Week 7: Preparing your final report & presentation		
Date and location	**Meeting minutes**	**Team members attended**
17/03/24, 17:00 - Zoom	Review the draft report and brainstorm any improvements to be implement before submission Discussed about the next week goals	Laurita Kunickaite, Om Sadigale, Faisal Qaderi, Yahya Habib,
Week 8: Live presentation & final report submission		
Date and location	**Meeting minutes**	**Team members attended**
24/03/26, 21:00 - Zoom	Preparing for the presentation tomorrow. Discussing the slides. Brainstorming the last changes for final report submission	Laurita Kunickaite, Om Sadigale, Faisal Qaderi, Yahya Habib, Ayaan Iqbal, Anas Kagigi

Background research

Water contamination is a growing global issue with significant implications for the environment, public health, and economic development (Fida et al., 2022). Water contaminants originate from various sources, including industrial discharge, agricultural runoff, urban sewage, and natural processes. Mechanical exercises such as fabricating, mining, and vitality generation discharge poisons such as overwhelming metals, chemicals, and poisons into water bodies. Agrarian hones including the utilisation of fertilizers, pesticides, and creature squander contribute to supplement contamination and pesticide runoff. Urbanization and populace development lead to an expanded wastewater era, contributing to microbial defilement and natural contamination in water sources. Moreover, common occasions such as disintegration, sedimentation, and volcanic movement can present contaminants in water bodies (UN, 2021).

Water contaminants include a wide extend of substances, including chemical, organic, and physical toxins. Chemical contaminants incorporate overwhelming metals, natural chemicals,

and mechanical toxins. Natural contaminants comprise pathogens such as microbes, infections, and parasites, which pose dangers to human well-being through waterborne illnesses. Physical contaminants incorporate dregs, flotsam and jetsam, and particulate matter, which can decrease water quality and oceanic environments (Rev, 2020).

Water contaminants have significant impacts on human well-being, biological systems, and socio-economic exercises. Sullied drinking water poses dangers of intense and unremitting well-being impacts. Waterborne maladies contribute to dismalness and mortality, especially in creating nations with a lack of sanitation foundation. Environmentally, water contaminants disturb oceanic environments, leading to territory corruption, biodiversity misfortune, and impeded water quality. Financially, water defilement comes about in costs related to healthcare, water treatment, natural remediation, and the misfortune of environmental administrations (UN,2023).

Overall, water contaminants posture noteworthy challenges to supportability, influencing human wellbeing, environments, and socio-economic well-being. Tending to water defilement requires facilitated endeavours to distinguish contamination sources, moderate impacts, and execute successful administration methodologies. By understanding the sources, sorts, impacts, and moderation measures of water contaminants, partners can work towards shielding water assets and advancing maintainable water administration. (Bashir et al.,2020).

Data set analysis

Data evaluation

Data evaluation was one of the key steps in the impact project. This process involves assessing the quality, accuracy, and relevance of the datasets obtained from the OECD.Stat.

The dataset provides information on the level of public equipment installed by countries to manage and abate water pollution. It shows the percentage of the national population connected to "public" sewerage networks and related treatment facilities, the percentage of the national population connected to "public" wastewater treatment plants, and the degree of treatment.

The main objectives were to identify inherent risks, biases, and gaps in the data, and assess the quality and reliability of six datasets to ensure that the subsequent analysis produces reliable and meaningful insights for fund managers.

Methodology:

1. Data identification
2. Quality assessment
3. Bias identification
4. Relevance analysis
5. Documentation and reporting

1. Generation and discharge of wastewater in volume:

- Source: Eurostat
- Data Source Credibility: Eurostat is in partnership with the European Statistical System (ESS) and is considered reliable with a legal obligation for trustworthiness. The main reason being is the partnership with ESS. It ensures the statistics provided for all the EU Member states are reliable. It is done by uniform criteria where they compare statistical data with different EU countries.
- Data Collection Method: Questionnaires are used annually, along with gap filling and corrections for earlier reference years.
- Bias: No bias was found.
- Conclusion: The data from this source is considered reliable.
- Links:
    1. Europa.eu. (2024). Available at: https://ec.europa.eu/eurostat/databrowser/view/env_ww_genv/default/table [Accessed 24 Mar. 2024].

    2. www.destatis.de. (n.d.). European Statistical System - German Federal Statistical Office. [online] Available at: https://www.destatis.de/Europa/EN/Methods/ESS/_inhalt.html#173226 [Accessed 27 Feb. 2024].

    3. ec.europa.eu. (n.d.). Eurostat and the European Statistical System. [online] Available at: https://ec.europa.eu/eurostat/statistics explained/index.php?title=Eurostat_and_the_European_Statistical_System#European_Statistical_System_.28ESS.29 [Accessed 27 Feb. 2024].

2. Mortality rate attributed to unsafe water, unsafe sanitation, and lack of hygiene:

- Source: The World Bank, World Health Organisation
- Data Source Credibility: Both The World Bank and the World Health Organisation are reputable sources. The world bank provided the governments of poor countries with loan to aid with capital projects. Many Researchers relay on their data for insights. The article by Jilian Clare Kohler and Andrea Bowra provides insight on how fair the data collection process is by implementing anti-corruption, transparency, and accountability.
- Data Collection Method: Death rates are calculated by dividing total deaths by the total population size, considering impacts like diarrhoeal diseases, intestinal nematode infection, and protein-energy malnutrition.
- Bias: Potential sampling bias as it does not cover impacts like typhoid, polio, etc.
- Conclusion: The data is considered reliable, but there might be some limitations due to the sampling bias.
- Links:

1. World Bank Gender Data Portal. (n.d.). Mortality rate attributed to unsafe water, unsafe sanitation and lack of hygiene (per 100,000 population). [online] Available at: https://genderdata.worldbank.org/indicators/sh-sta-wash-p5/?gender=total.

2. Kohler, J.C. and Bowra, A. (2020). Exploring anti-corruption, transparency, and accountability in the World Health Organization, the United Nations Development Programme, the World Bank Group, and the Global Fund to Fight AIDS, Tuberculosis and Malaria. Globalization and Health, 16(1). doi:https://doi.org/10.1186/s12992-020-00629-5.

3. Premature deaths due to UNSAFE WASH:

- Source: OCED.Stat
- Data Source Credibility: OCED.Stat is considered reputable, and data is based on the Global Burden of Diseases (GBD) from articles in 'The Lancet.' As they considered these many aspects of the data as there are less chances of data being biased hence this data can considered reliable
- Data Collection Method: Examining trends from 1990 to the present, considering 204 countries, 369 diseases and injuries, and 87 risk factors.
- Bias: No bias was found.
- Conclusion: The source is reliable.
- Links:

1. stats.oecd.org. (n.d.). Mortality, morbidity and welfare cost from exposure to environment-related risks. [online] Available at: https://stats.oecd.org/Index.aspx?DataSetCode=EXP_MORSC.

2. www.thelancet.com. (n.d.). About the Global Burden of Disease. [online] Available at: https://www.thelancet.com/gbd/about.

4. Proportion of bodies of water with good ambient water quality:

- Source: UNWater
- Data Source Credibility: UNWater is a monitoring hub and provides reliable data.
- Data Collection Method: In-situ measurements of water quality parameters, comparing measured values to national target levels. They emphasize on water related issues/ measures such as Water quality, wastewater management, water scarcity and they acknowledge the value of water and make efforts to improve the water value. It is a reputable source when considering the field of water sanitation.
- Bias: No biases were found, but data provided may not be sufficient.
- Conclusion: The data is considered reliable, but there might be limitations due to insufficient data.
- Links:

1. sdg6data.org. (n.d.). Indicator | SDG 6 Data. [online] Available at:

https://sdg6data.org/en/indicator/6.3.2.

2. United Nations (2023). Water Quality and Wastewater. [online] UN-Water. Available at: https://www.unwater.org/water-facts/water-quality-and-wastewater.

3. United Nations (2021). Water Scarcity. [online] UN-Water. Available at: https://www.unwater.org/water-facts/water-scarcity.

4. UN-Water. (n.d.). UN World Water Development Report 2021. [online] Available at: https://www.unwater.org/publications/un-world-water-development-report-2021.

5. Wastewater Discharges Per Year Per Country:

- Source: OCED.stat
- Data Source Credibility: OCED.stat is a reputable source, and the data collection method is guided by UN-Habitat, WHO, UNSD, Eurostat, and UN-Water. As mentioned earlier, the world bank provided the governments of poor countries with loan to aid with capital projects. Many Researchers relay on their data for insights. The article by Jilian Clare Kohler and Andrea Bowra provides insight on how fair the data collection process is by implementing anti-corruption, transparency, and accountability.
- Data Collection Method: The UN-Habitat, WHO, UNSD for monitoring methodology. Eurostat for Questionnaire for EU countries. UNSD Questionnaire for OCED countries. UN-Water for Data collection process and timeline. Collaboration among various organizations for collecting wastewater discharge data.
- Bias: No biases were found.
- Conclusion: The data from this source is considered reliable.
- Links:

1. stats.oecd.org. (n.d.). Generation and discharge of wastewater. [online] Available at: https://stats.oecd.org/Index.aspx?DataSetCode=WATER_DISCHARGE.

2. Kohler, J.C. and Bowra, A. (2020). Exploring anti-corruption, transparency, and accountability in the World Health Organization, the United Nations Development Programme, the World Bank Group, and the Global Fund to Fight AIDS, Tuberculosis and Malaria. Globalization and Health, 16(1). doi:https://doi.org/10.1186/s12992-020-00629-5.

6. Annual freshwater withdrawals:

- Source: Our World in Data, The World Bank, WHO/ UNICEF JMP, Food Agriculture Organization of the United Nations
- Data Source Credibility: The sources are credible as they mention clearly the methodologies used and it is done by professionals. The founder of Our World in

Data Max Roser is a Programme director of Oxford Martin Programme on Global Development. For data collection they have considered various aspects such as Industry and Agriculture.

- Data Collection Method: Annual freshwater withdrawals encompass total water extraction, excluding evaporation losses from storage basins. The dataset includes water from desalination plants, especially in countries where they are a significant source. Water withdrawals may exceed 100 percent of total renewable resources in cases of substantial extraction from non-renewable aquifers, desalination plants, or significant water reuse. Withdrawals for agriculture and industry cover irrigation, livestock production, and direct industrial use, including cooling for thermoelectric plants. Domestic uses include withdrawals for drinking water, municipal supply, public services, commercial establishments, and homes.
- Bias: No Bias found
- Conclusion: The Data is Reliable.
- Links:

1. Our World in Data. (n.d.). Annual freshwater withdrawals. [online] Available at: https://ourworldindata.org/grapher/annual-freshwater-withdrawals?tab=table [Accessed 27 Feb. 2024].

2. Media Bias/Fact Check. (2017). Our World In Data - Media Bias/Fact Check. [online] Available at: https://mediabiasfactcheck.com/our-world-in-data/.

7. Death rate from Unsafe Water sources

- Source: Our World in Data, IHME, Global Burden of Disease (2019)
- Data Source Credibility: The credibility of the sources is established through transparency regarding the methodologies employed, managed by skilled professionals. IHME has engaged in global collaborations with numerous organizations, including the World Health Organization (WHO), with whom they have established a memorandum of understanding.
- Data Collection Method: To calculate death and DALY rates, sum up the deaths and DALYs for each country in a region of interest. Then, divide these totals by the corresponding regional population. This gives the death and DALY rates for that specific region.
- Bias: No bias found
- Conclusion: The dataset is reliable
- Links:

1. Our World in Data. (n.d.). *Death rates from unsafe water sources*. [online] Available at: https://ourworldindata.org/grapher/death-rates-unsafe-water.

2. www.who.int. (n.d.). *WHO/Europe and IHME sign agreement cementing collaboration on forecasting of health data*. [online] Available at:

https://www.who.int/europe/news/item/07-02-2022-who-europe-and-ihme-sign-agreement-cementing-collaboration-on-forecasting-of-health-data#:~:text=WHO%2FEurope%20and%20the%20Institute%20for%20Health%20Metrics%20and [Accessed 24 Mar. 2024].

3. www.who.int. (n.d.). *Indicator Metadata Registry Details*. [online] Available at: https://www.who.int/data/gho/indicator-metadata-registry/imr-details/2260.

8. Water, Sanitation and Hygiene (WASH) Data Explorer:

- Source: Our World in Data, WHO/UNICEF Joint Monitoring Programme (JMP) for Water Supply and Sanitation
- Data Source Credibility: WHO and UNICEF, as specialized United Nations agencies, possess extensive expertise in their designated fields. While WHO concentrates on addressing worldwide health concerns. UNICEF's focus lies in safeguarding children's rights, offering aid and advocacy across domains like education, nutrition, and safeguarding against exploitation. Their considerable expertise, coupled with the transparency of their research, amplifies their accountability.
- Data Collection Method: They collect data from national governments through surveys which gather information on Drinking water, sanitation and hygiene. There are various estimation methods used one of them being drawing a best fit line of the data points generated from rural, urban and national estimates for each country.
- Bias: No Bias found
- Conclusion: The dataset is reliable
- Links:

1. www.unicef.org. (n.d.). *For every child, results | UNICEF*. [online] Available at: https://www.unicef.org/results#:~:text=UNICEF%20is%20the [Accessed 24 Mar. 2024].

2. World Health Organization (n.d.). *Credible and trusted*. [online] www.who.int. Available at: https://www.who.int/about/communications/credible-and-trusted.

3. washdata.org. (n.d.). *Estimation methods | JMP*. [online] Available at:https://washdata.org/monitoring/methods/estimation-methods.

Data pre-processing

Raw data overview

Before implementing data preprocessing, the raw data is depicted in its original form as presented below for subsequent processing.

GitHub link: https://github.com/YahyaHabib/-Water---contaminants-and-levels (reference for code script)

Datafile 1: "annual-freshwater-withdrawals.csv" consists of the following characteristics:

- Number of Rows: 5827
- Number of Columns: 4

The columns data type in this dataset are (figure 1):

- **Entity** (Categorical): Refers to the country or region name.
- **Code** (Categorical): The ISO code or similar identifier for the entity.
- **Year** (Numerical): The year of the observation.
- **Annual freshwater withdrawals, total (billion cubic meters)** (Numerical): The total annual freshwater withdrawals in billion cubic meters.

```
[ ]  freshwater_withdrawals_df.dtypes

Entity                                                  object
Code                                                    object
Year                                                     int64
Annual freshwater withdrawals, total (billion cubic meters)    float64
dtype: object
```

Figure 1

Datafile 2: "Wastewater Discharges Per Year Per Country.xlsx" consists of the following characteristics:

- Number of Rows: 205
- Number of Columns: 14

The columns data type in this dataset are:

- **Country** (Categorical): Refers to the country name.
- **Year** (Numerical): The year of the observation.
- **Total discharges to Inland waters(million m3)** (Numerical): Total wastewater discharges to inland waters.
- **Total discharges to the sea(million m3)** (Numerical): Total wastewater discharges to the sea.
- **Agricultural (incl. forestry + fisheries) wastewater, all sources, direct discharges(million m3)** (Numerical): Wastewater from agriculture, forestry, and fisheries discharged directly.
- **Urban wastewater, all sources, discharged without treatment(million m3)** (Numerical): Urban wastewater discharged without treatment.
- **Industrial wastewater, all sources, discharged without treatment(million m3)** (Numerical): Industrial wastewater discharged without treatment.

Datafile 3: “Premature_deaths_due_to_UNSAFE_WASH.xlsx” consists of the following characteristics:

- Number of Rows: 1206
- Number of Columns: 5

The columns data type in this dataset are:

- **Country** (Categorical): Refers to the country name.
- **Year** (Numerical): The year of the observation.
- **Premature_Death_Count** (Numerical): The count of premature deaths due to UNSAFE WASH activities.
- **Risk** (Categorical): The type of risk associated with the deaths.
- **Health_Impact** (Categorical): The impact level of health due to the deaths categorized into levels (e.g., Low, Medium, High)

Datafile 4: "Mortality rate attributed to unsafe water, unsafe sanitation and lack of hygiene (per 100,000 population)" consists of the following characteristics:

- Number of Rows: 266
- Number of Columns: 35
-

The columns data type in this dataset are (figure 2):

- **Country Name** (Categorical): Refers to the country or region name.
- **Country Code** (Categorical): The ISO code or similar identifier for the entity.
- **Series Name** (Categorical): The name of the data series.
- **Series Code** (Categorical): The code for the data series.
- **1990 [YR1990] to 2020 [YR2020]** (Categorical): The mortality rate attributed to unsafe water, unsafe sanitation, and lack of hygiene for each year from 1990 to 2020. The data for these years are initially categorized as categorical, which may include non-numeric values or notes.

```
# Display the first few rows of the dataset and its info to understand its structure
death_rate_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6840 entries, 0 to 6839
Data columns (total 4 columns):
 #   Column                                                                                                  Non-Null Count  Dtype
---  ------                                                                                                  --------------  -----
 0   Entity                                                                                                  6840 non-null   object
 1   Code                                                                                                    6150 non-null   object
 2   Year                                                                                                    6840 non-null   int64
 3   Deaths that are from all causes attributed to unsafe water source per 100,000 people, in both sexes aged age-standardized  6840 non-null   float64
dtypes: float64(1), int64(1), object(2)
memory usage: 213.9+ KB
```

Figure 2

Datafile 5: “Proportion of bodies of water with good ambient water quality.xlsx” consists of the following characteristics:

- Number of Rows: 244
- Number of Columns: 67

The columns data type in this dataset are:

- **Country Name** (Categorical): Refers to the country name.
- **Country Code** (Categorical): The ISO code or similar identifier for the country.
- **Series Name** (Categorical): Name of the data series.
- **Series Code** (Categorical): Code for the data series.
- **1960 [YR1960] to 2022 [YR2022]** (Categorical): Entries for each year, classified as categorical due to the presence of non-numeric entries, indicating proportions or data related to the quality of water bodies.

Datafile 6: “Generation and discharge of wastewater in volume.xlsx” consists of the following characteristics:

- Summary Sheet: provides metadata including descriptions, data update logs, and context for the dataset.

- Structure Sheet: outlines the dataset's dimensions, such as categories of wastewater generation, treatment, and discharge. It details the data's categorization by source and destination, emphasizing the annual frequency and specifying the years covered.

General Observations:

- The datafile is structured across 34 sheets, each labelled "Sheet 1" through "Sheet 34," encompassing various facets of wastewater generation, treatment, and discharge. The data span different categories, including overall sources, point sources, and detailed segments for industrial and agricultural wastewater contributions.

- The datasets demonstrate inconsistency in terms of geographic and temporal coverage. Certain countries are represented in some sheets but absent in others, indicating variability in data availability or collection practices across regions. Additionally, the extent of historical data varies significantly among countries, with some having extensive yearly records, while others have sparse data points. This inconsistency highlights the diverse nature of wastewater management data collection and reporting standards globally, affecting the comprehensiveness and comparability of the datasets.

Data relevance and selection

After conducting data evaluation and checking for missing values, the decision was made to use these specific datasets for further data pre-processing and analysis:

- **Annual Freshwater Withdrawals:** the dataset was chosen because it has most of the data for many countries and not much missing information.

- **Premature deaths due to UNSAFE_WASH**: the dataset has been for further analysis due to its commendable attributes, featuring restricted gaps in data and widespread availability across numerous countries. This dataset not only contributes valuable insights but also aligns with the common countries prevalent throughout the document.

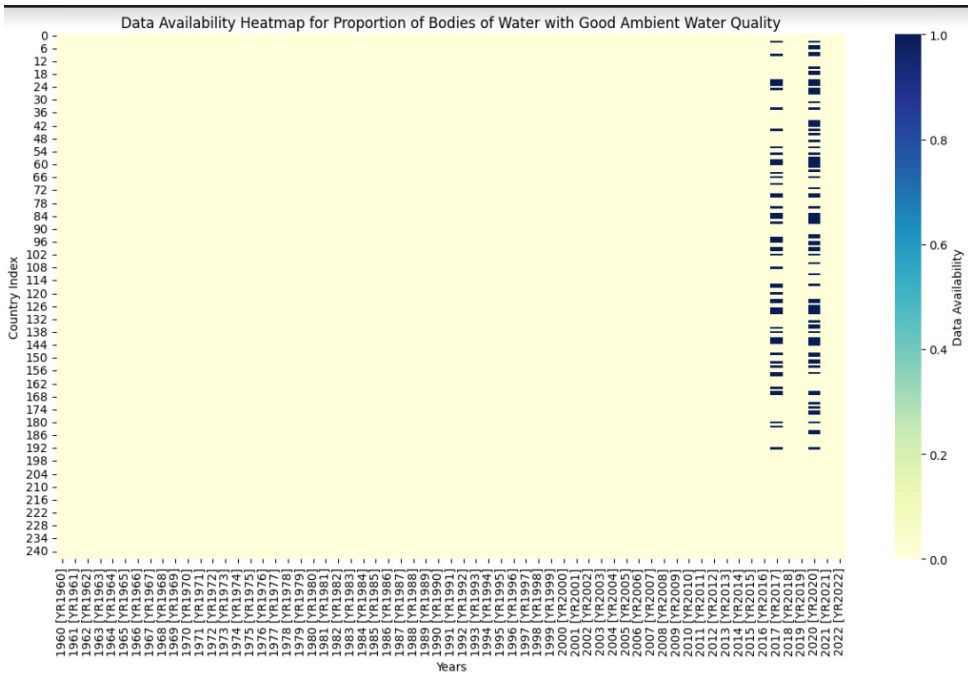
- **Generation and discharge of wastewater in volume:** it had a lot of relevant information for global metric, which was used for the regression model.

- **Wastewater Discharges Per Year Per Country**: this dataset is crucial for further analysis. It has good coverage, not much missing data, and contains many metrics such as waste water of agriculture, industry waste water and etc. Also, common countries can be extracted from this data set, which have correlation with other data sets.

Data sets have not been used:

- **Proportion of bodies of water with good ambient water quality:** the dataset was not chosen due to its inadequacy regarding the specified years and countries covered in the document. The heatmap below represents missing values, evidencing why this data set was not used.

- **Mortality rate**



attributed to unsafe water, unsafe sanitation, and lack of hygiene (per 100,000 population): this particular dataset was left out because it had limited and not accurate data. It only covered one country, and there were many missing values within it.

Applied methodology in data pre-processing

Data pre-processing is a process which involves cleaning and transforming raw data into a tidy format for further analysis and visualisation. It includes several steps such as data exploration, handling missing data: Identify missing values across datasets. Apply imputation techniques where reasonable (e.g., median imputation for numerical variables, mode imputation for categorical variables) or consider removing observations with excessive missing information.

Also, removing duplicates: checking for and removing any duplicate records across datasets, ensuring each observation is unique and correctly represented. Moreover, correcting errors: performing a thorough review of data for inconsistencies or anomalies (e.g., negative values where only positive values make sense, unlikely outliers) and correcting these errors based on contextual knowledge or by consulting additional sources.

By using Python programming language these steps were applied to valuated data sets:

Annual Freshwater Withdrawals:

The initial examination revealed 78 missing observations in the "Code" section. A deeper exploration highlighted that the "Entity" column included not only country names but also aggregated entities like "Middle East," "Middle-Income," and "Lower-Income," which do not possess ISO codes. To preserve the integrity and utility of the dataset while maintaining inclusivity of all entities, it was opted to replace missing codes with "N/A" rather than removing these observations. This approach ensured that it was retained valuable aggregated data for comprehensive insights. Post-adjustment, the dataset was devoid of any duplicate or missing data, and the data types were verified to be appropriate for further analytical needs (figure 3).

```
[ ] freshwater_withdrawals_df.isnull().sum()

Entity                                                  0
Code                                                   78
Year                                                    0
Annual freshwater withdrawals, total (billion cubic meters)    0
dtype: int64
```

```
missing_code_rows = freshwater_withdrawals_df[freshwater_withdrawals_df['Code'].isnull()]

missing_code_rows.sample(20)
```

	Entity	Code	Year	Annual freshwater withdrawals, total (billion cubic meters)
3505	Middle-income countries	NaN	2015	2.828180e+12
1729	Europe and Central Asia (WB)	NaN	2014	4.857940e+11
3941	North America (WB)	NaN	2016	4.803880e+11
3116	Low-income countries	NaN	2018	1.196610e+11
3506	Middle-income countries	NaN	2016	2.858490e+12
3120	Lower-middle-income countries	NaN	2016	1.643930e+12

Figure 3

Premature Deaths:

This dataset was found to be complete with no missing or duplicated data upon initial assessment. Furthermore, the data types were correctly formatted, indicating that the dataset was well-prepared for immediate analysis without the need for additional cleaning steps (Figure 4)

```
premature_deaths_df.isnull().sum()

Country                0
Year                   0
Premature_Death_Count  0
Risk                   0
Health_Impact          0
dtype: int64
```

```
[ ] premature_deaths_df.duplicated().sum()

0
```

Figure 4

Wastewater Discharges:

Similar to the "Premature Deaths" dataset, the "Wastewater Discharges" dataset required no additional cleaning. It was devoid of missing values and duplicates, with all data types being correctly formatted. This readiness significantly streamlined the pre-processing phase, allowing for a more efficient transition to the analysis stage (Figure 5)

```
[ ] wastewater_data.isnull().sum()

Country                                                                                    0
Year                                                                                       0
Total discharges to Inland waters(million m3)                                              0
Total discharges to the sea(million m3)                                                    0
Agricultural (incl. forestry + fisheries) wastewater, all sources, direct discharges(million m3)    0
Urban wastewater, all sources, discharged without treatment(million m3)                    0
Industrial wastewater, all sources, discharged without treatment(million m3)               0
dtype: int64

[ ] wastewater_data.duplicated().sum()

0
```

Figure 5

Water Quality:

The "Water Quality" dataset presented unique challenges, primarily due to its wide format. It was transformed it into a long format to address the issue of many years lacking data points, thereby enhancing the dataset's usability for temporal analysis. In the process of transformation, duplicate rows were identified and subsequently removed to ensure data integrity. A peculiar discovery was the recording of the "Proportion of Good Water Quality" as less than or equal to 0 for two countries. Recognizing this as a data entry error or a lack of recorded data, these errors were removed to maintain the dataset's accuracy (refer to code script).

Each dataset underwent a meticulous cleaning process tailored to its unique characteristics and the challenges it presented. These pre-processing steps were essential in building a robust foundation for our analysis, ensuring that findings would be based on reliable and accurate data.

Data Integration Process

After cleaning datasets, it was focused on combining them to analyse water issues globally and by country. Integration of “Annual Freshwater Withdrawals” and “Premature Deaths” datasets was achieved through a strategic merging process, utilising Python. This process entailed (refer to code script):

- **Harmonising Key Identifiers:** "Country" and "Year" columns across both datasets were standardized to serve as common keys for merging. This included renaming columns for consistency and ensuring data types matched (e.g., country names and years were formatted identically).

- **Merging Data**: using Python's pandas’ library, it was executed an inner join operation on the "Country" and "Year" columns. This method ensured that only records with matching countries and years were combined, providing a dataset that

contained both water withdrawal volumes and premature death counts for each country and year available.

- **Assessing and Addressing Data Completeness:** post-merger, the integrated dataset were evaluated for any anomalies or significant data losses. This step was crucial in ensuring the merged dataset was comprehensive and reflective of the temporal and geographic scopes of our study.

- **Global and Country-Specific Analysis:** with the integrated dataset, it needed to be positioned to analyse water-related health impacts both globally and by individual countries. This facilitated a nuanced understanding of how freshwater withdrawals might correlate with health outcomes across different regions and over time.

Challenges with Further Dataset Integration

While the merger of the "Annual Freshwater Withdrawals" and "Premature Deaths" datasets were fruitful, integrating additional datasets, such as "Wastewater Discharges" and "Water Quality," presented notable challenges:

- Temporal Variability: Some datasets exhibited limited temporal coverage or inconsistent year ranges, complicating the alignment with our core integrated dataset which was structured to facilitate a longitudinal analysis.

- Country-Specific Data Availability: The granularity and availability of data varied significantly by country in the additional datasets. This inconsistency posed a challenge to achieving a comprehensive global analysis, as integrating these datasets could result in substantial data loss or biased insights favouring regions with more complete data.

To navigate these challenges, it was prioritized maintaining the integrity and usability of integrated dataset, opting to proceed with a focused analysis on the successfully merged "Annual Freshwater Withdrawals" and "Premature Deaths" data. While this decision meant forgoing the direct integration of all datasets, it ensured analysis remained robust, accurate, and globally relevant. Future efforts may explore methodologies for incorporating additional datasets, potentially leveraging more sophisticated data imputation techniques or seeking out supplementary data sources to fill existing gaps.

Statistical modelling and error margins

In this project, statistical modelling techniques were implemented, such as regression analysis, error margin calculation, and projection techniques, which is essential for ensuring the accuracy and reliability of data in different lenses.

Applied, regression analysis allowed to identify relationships between variables, aiding in understanding trends and making predictions. Meanwhile, error margin calculation helped assess the precision of data measurements, identifying potential biases or uncertainties. Moreover, the implemented projection formula (as per below) enables the extrapolation of insights from existing data to forecast future trends/ impacts. By integrating these statistical methods, the team derived meaningful insights, made informed decisions, and mitigated risks effectively. Most importantly, it played a crucial role in enhancing the quality and credibility of analytical results, thus contributing to evidence-based decision-making for Koru Impact solution success

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Here's what each term represents:

Y is the dependent variable (the variable being predicted).	β_0 is the intercept of the regression line (the value of Y when X is 0).	β_1 is the slope of the regression line (how much Y changes for a one-unit change in X).	X is the independent variable (the variable used to predict Y).	ϵ is the error term (the difference between the observed values and the values predicted by the model).

Firstly, in order to find relationships between different metrics correlation heatmap was created. Correlation values range from -1 to 1, where 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and values close to 0 suggest little to no linear relationship. Moreover, reveals how each predictor correlates with the others, including potential relationships that could inform further analysis. For instance, high positive correlations between different types of wastewater discharges could indicate common underlying factors or effects (figure 11)

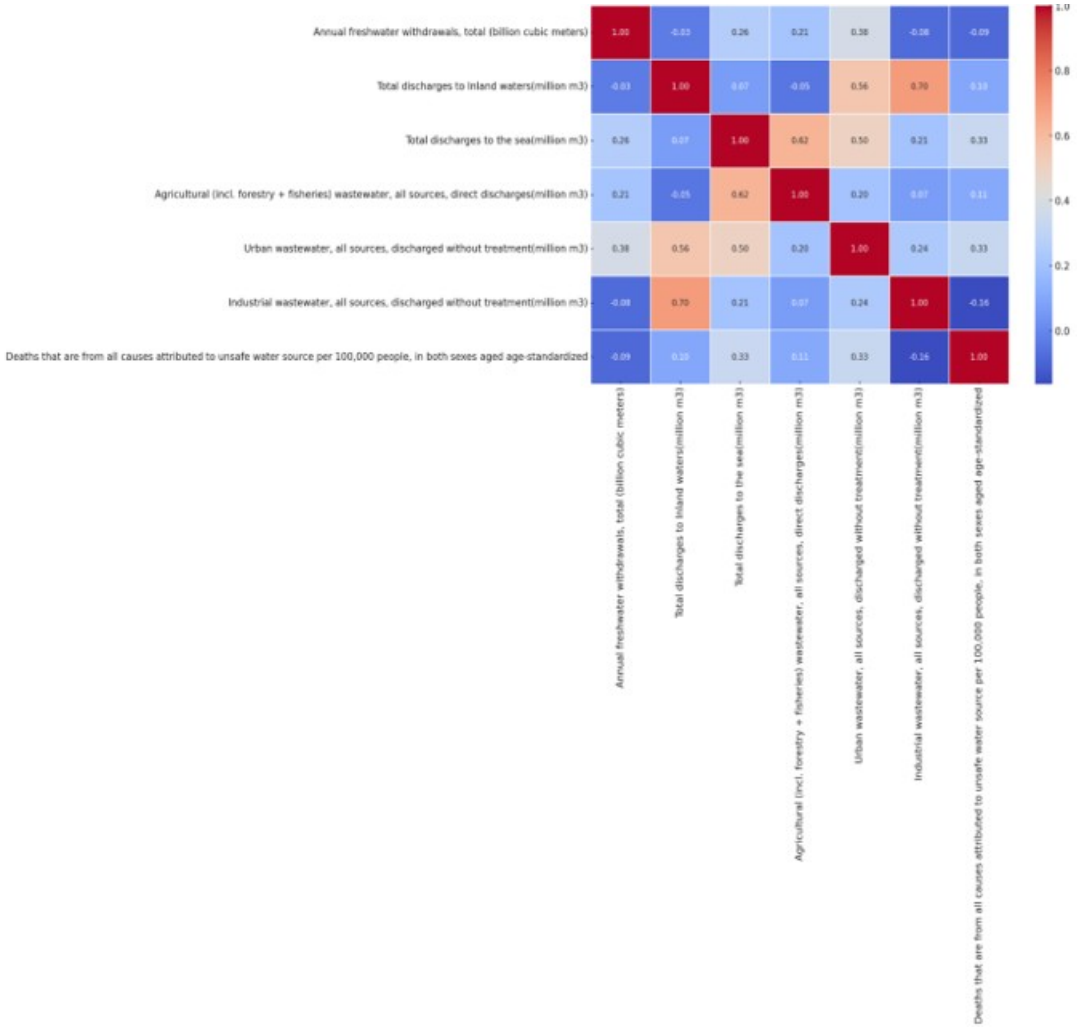


Figure 11

As it can be seen from the heatmap, the strongest correlation is between industrial water discharges without any treatment and total discharges to inland waters.
All-inclusive, the correlation analysis provides a foundational understanding of the relationships between various environmental and infrastructural factors and deaths due to unsafe WASH practices. This insight can guide more detailed statistical modelling and inform policy and intervention strategies aimed at reducing such deaths.

Error Margins

It has computed both the mean values and their respective error margins for each key metric. The margin of error is a crucial statistical measure that helps us understand the confidence interval within which the true mean of our population is likely to fall. The confidence level is established at 95%, denoting a 95% certainty regarding the actual mean's inclusion within this specified range. Hereafter, a comprehensive elucidation is furnished delineating the implications of these error margins for each metric.

1. Deaths Attributed to Unsafe Water Sources (per 100,000 people, age-standardized for both sexes)

- **Mean**: 0.236 deaths
- **Margin of Error**: ±0.102 deaths
- **Implication**: 95% confident level represents that the true mean of deaths per 100,000 people due to unsafe water sources lies between 0.134 and 0.338. This metric highlights critical health risks associated with water safety and underscores the importance of improving water quality.

2. Premature Death Count

- **Mean**: 67.69 deaths
- **Margin of Error**: ±14.32 deaths
- **Implication**: The true mean of premature deaths in our dataset is estimated to be within the range of 53.37 and 82.01, with a 95% confidence level. This range quantifies the health impact of environmental factors and supports targeted interventions.

3. Annual Freshwater Withdrawals (Total in billion cubic meters)

- **Mean**: 3,280,924,720 m^3
- **Margin of Error**: ±916,946,805.74 m^3
- **Implication**: There is a significant variability in freshwater withdrawals, indicating varying levels of water resource management and consumption patterns. The wide margin suggests the need for sustainable water use policies.

4. Total Discharges to Inland Waters (in million m^3)

- **Mean**: 862.19 m^3
- **Margin of Error**: ±228.68 m^3
- **Implication**: With a 95% confidence interval, the volume of discharges to inland waters is estimated to be between 633.51 and 1,090.87 million m^3. This emphasizes the environmental impact of waste disposal and the necessity for effective waste management practices.

5. Total Discharges to the Sea (in million m^3)

- **Mean**: 570.82 m^3
- **Margin of Error**: ±187.23 m^3
- **Implication**: The true mean discharge volume to the sea is likely between 383.59 and 758.05 million m^3, indicating the maritime impact of waste discharges. It calls for marine protection measures to preserve aquatic ecosystems.

6. Agricultural Wastewater, Direct Discharges (in million m^3)

- **Mean**: 121.25 m^3
- **Margin of Error**: ±50.55 m^3
- **Implication**: The agricultural sector's impact on water quality is highlighted, with a true mean discharge volume estimated between 70.70 and 171.80 million m^3. This points to the need for sustainable agricultural practices.

7. Urban Wastewater, Discharged Without Treatment (in million m^3)

- **Mean**: 93.32 m^3
- **Margin of Error**: ±12.59 m^3

- **Implication**: The relatively narrow margin of error indicates a more consistent measure of untreated urban wastewater. Efforts to improve urban wastewater treatment could significantly benefit public health and environmental quality.

8. Industrial Wastewater, Discharged Without Treatment (in million m^3)

- **Mean**: 175.61 m^3
- **Margin of Error**: ±96.45 m^3
- **Implication**: The wide margin of error suggests considerable variability in industrial wastewater management, underscoring the need for stricter regulations and cleaner production technologies.

Analysis and recommendations

Wastewater discharges impact on Health

After applying a correlation matrix, it was discovered that wastewater from various sources exhibits the strongest correlation with other metrics. Consequently, it was selected as a global metric for gaining insights into this project, owing to its significant impact on health, environment, carbon density and society.

Urban Wastewater's Impact on Health:

- **Correlation:** it was discovered a significant relationship between urban wastewater discharge and premature death rates. Our regression model demonstrates that increases in urban wastewater discharge are associated with rises in premature death rates.
- **Quantified Impact:** the analysis yields a positive coefficient, indicating a direct impact on health. Specifically, for every increase of 1 million cubic meters in urban wastewater discharge, the premature death count rises by approximately 0.0261. This finding highlights the health risks associated with inadequate wastewater management (figure 12).

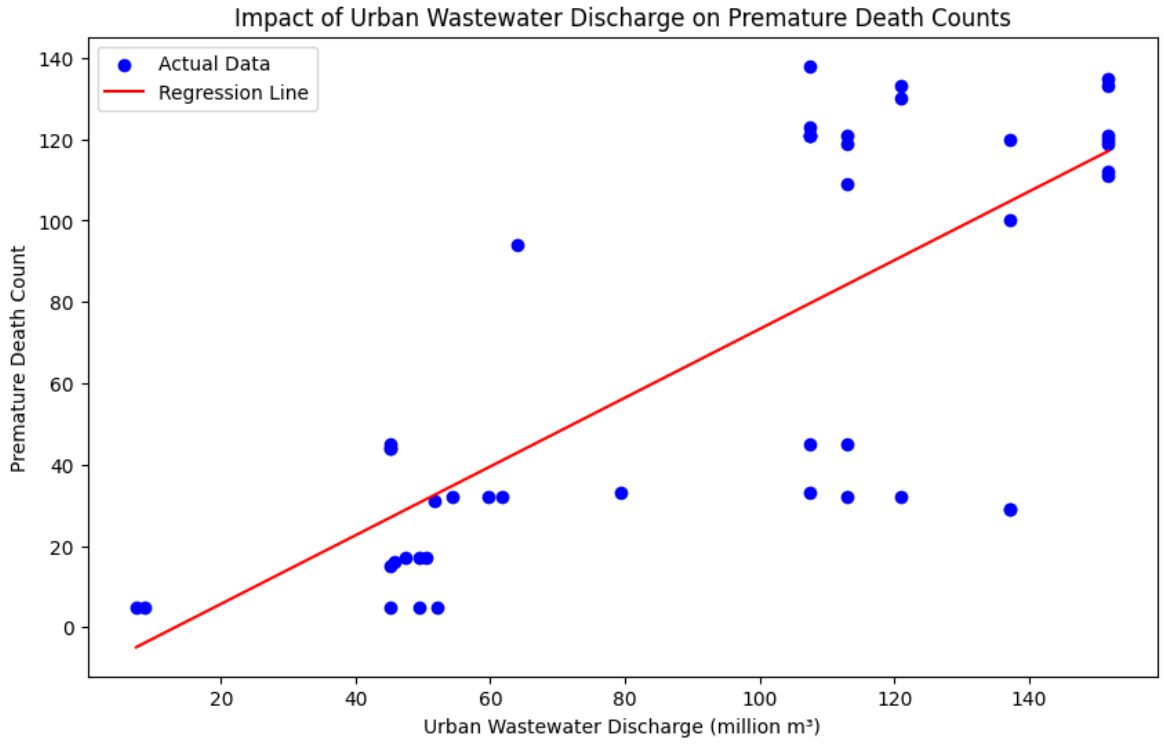


Figure 12

Model Parameters Explained

- **Coefficient (0.0261):** This figure represents the change in the number of premature deaths for each additional million cubic meters of urban wastewater discharged. It quantifies the direct health risks posed by wastewater, highlighting the relationship between environmental management and public health.
- **Intercept (45.1792):** The intercept offers an estimate of the base premature death count, assuming no urban wastewater discharge. This value provides a starting point for understanding the model's predictions in the context of an ideal scenario with no wastewater impact.

Linear Regression Graph Insights:

- The graph plotting urban wastewater discharge against premature death count illustrates the predictive relationship established by our regression model. The actual data points (blue dots) and the regression line (red line) together highlight how increases in wastewater discharge correlate with higher premature death counts. This visual representation reinforces our findings, emphasizing the need for effective wastewater management.

Premature Death Count/Global Health Metric Formula

- Analysis closes in a formula that quantifies the health impact of urban wastewater discharge on a global scale, facilitating the assessment of environmental policies or investments. The formula, incorporating the model's coefficient and intercept, offers

a practical tool for estimating premature death counts based on wastewater discharge volumes.

- **Estimated Premature Death Count/ Global Health Metric Formula** = (Urban Wastewater Discharge in million m^3 × 0.0261) + 45.1792 ± RMSE

Residuals vs Fitted Values and RMSE:

- **Residuals vs Fitted Values:** The plot of residuals (the differences between actual and predicted values) against fitted values helps us assess the model's accuracy and the appropriateness of its assumptions. From this it could observe there is a random scatter, which means our model seems to fit well. The random scatter of residuals suggests both linearity in the relationship and homoscedasticity (constant error variance). This implies unbiased model predictions across the range of wastewater discharge values (figure 13)

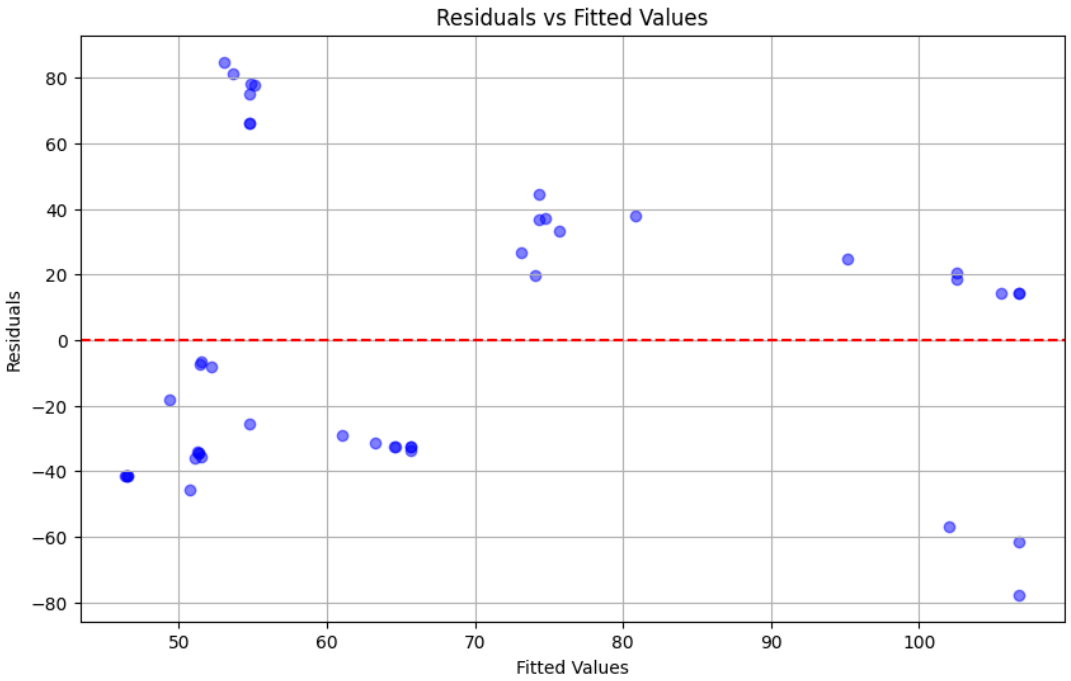


Figure 13

- **RMSE (Root Mean Squared Error) ~44.07:** This metric provides an average deviation of our model's predictions from the actual premature death counts, serving as an error margin. It quantifies the model's prediction accuracy, with lower values indicating better performance. RMSE suggests that, on average, the model's predictions are about 44 premature deaths off from the actual numbers, highlighting the potential variability in predictions.

Wastewater discharges impact on Environment

Agricultural Wastewater discharge's impact on sea:

As per correlation matrix, the correlation was found between agricultural wastewater discharges and total discharges to the sea. Hence, the regression visualisation indicates this correlation. The positive slope suggests that as the discharge of the volume of agricultural wastewater increases, the total discharges to the sea increases respectively

Applied progression formula (Agricultural Wastewater Discharges = 0.1682 × Total Discharges to the Sea (million m^3) + 25.2254) indicates that for every additional million cubic meters of agricultural wastewater discharges, it increases total discharge to the sea by an average of 0.1682 (million m^3).

The visualization provides a scatter plot of the actual data points (blue) and the predicted regression line (red). The shaded area represents the 95% confidence interval for the regression predictions, providing a range where we expect the actual value to lie with 95% confidence (figure 14).

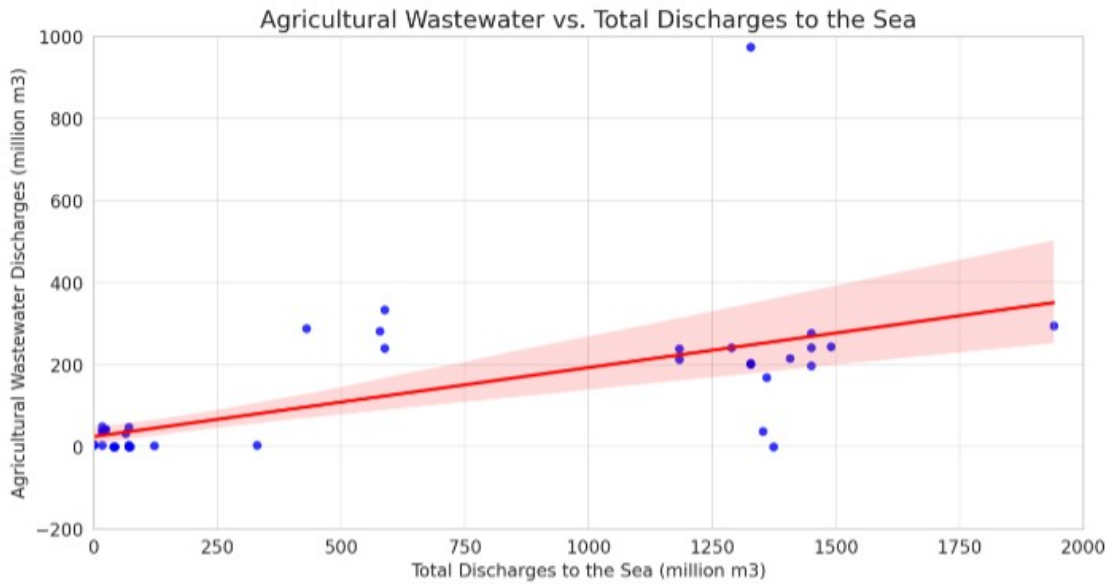


Figure 14

The model's coefficient of determination suggests that approximately 38.82% of the variability in agricultural wastewater discharges can be explained by the total discharges to the sea. The confidence intervals are relatively tight around the regression line, indicating a certain degree of precision in the model's predictions.

Agricultural Wastewater discharge's impact on inland waters:

Opposite to the impact on the sea, it was found a negative corelation to between agricultural wastewater discharges and total discharges to the inland waters. This implies that an increase in inland discharges is associated with a slight decrease in agricultural wastewater discharges. This relationship could be influenced by various factors, including

the management practices for agricultural and industrial wastewater or the geographical distribution of agricultural areas relative to inland water bodies.

Applied progression formula: Agricultural wastewater discharges = −0.01035 × Total discharges to Inland waters (million m^3) + 130.1775 Agricultural wastewater discharges = −0.01035 × Total discharges to Inland waters (million m^3) + 130.1775 indicates that for every million cubic meters increase in inland discharges, the volume of agricultural wastewater discharges decreases by approximately 0.01035 million cubic meters.

These models quantify the association between the total discharges and agricultural wastewater discharges, with positive association for sea discharges and a slight negative association for inland discharges (figure 15)

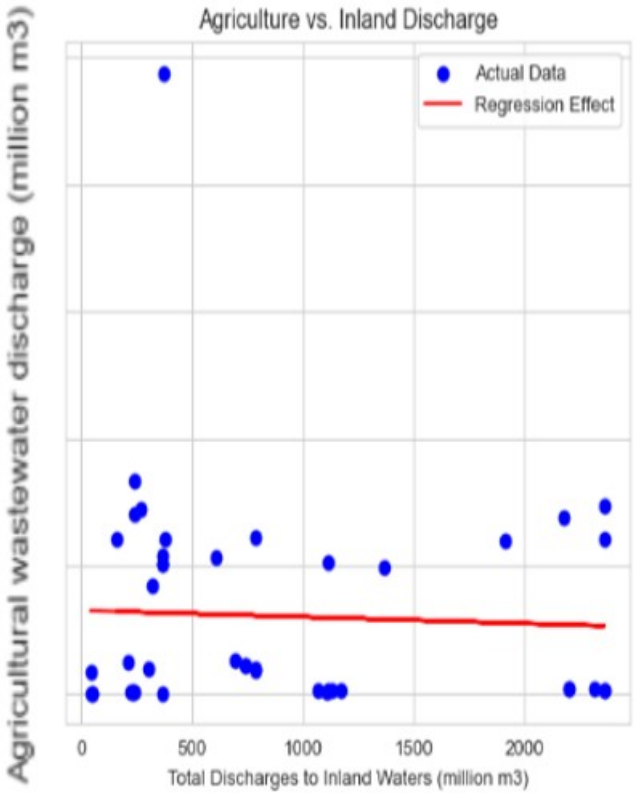


Figure 15

Urban Wastewater discharge's impact on sea:

Applied progression model (Urban Wastewater Discharges = 0.03337 × Total Discharges to the Sea (million m^3) + 74.2727) suggests that for every million cubic meters increase urban wastewater discharges, the volume of total discharges to sea increases by approximately 0.03337 million cubic meters.
The coefficient of determination, R^2, for this regression is approximately 0.2462, which indicates that around 24.62% of the variability in urban wastewater discharges can be explained by the total discharge to the sea. It's a modest fit, indicating that while there is some relationship, a significant portion of the variability is not captured by this model, and another factor may be influencing urban wastewater discharges (figure 16).

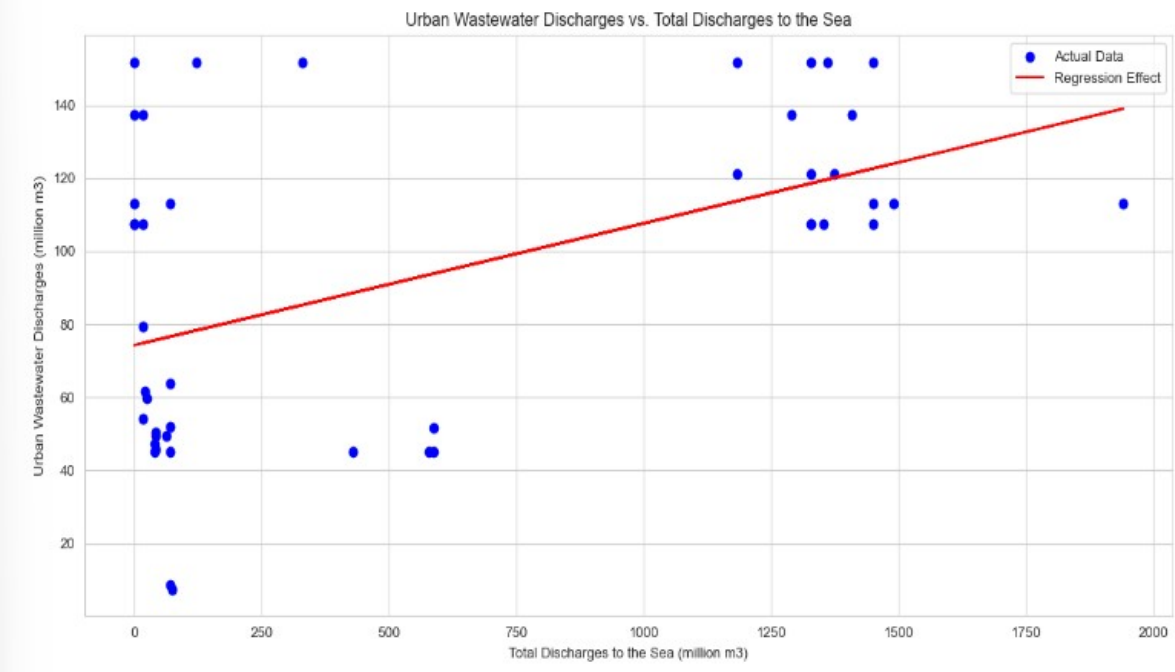


Figure 16

Industrial Wastewater discharge's impact on inland waters:

The regression analysis demonstrates a positive relationship between the volume of industrial wastewater discharges and the volume of inland water discharges. The model's moderate R^2 value indicates a substantial association, yet it also suggests that other factors play a role in determining industrial wastewater discharges that the model does not capture.
Applied progression formula: Industrial Wastewater Discharges = 0.2957 × Total Discharges to Inland Waters (million m^3) − 79.3546 where coefficient 0.2957 suggests that for each million cubic meters increase in the industrial wastewater discharges, inland water discharges increase by approximately 0.2957 million cubic meters. The R^2 value is 0.4916, indicating that about 49.16% of the variability in industrial wastewater discharges is explained by the model, showing a moderate fit (figure 17)

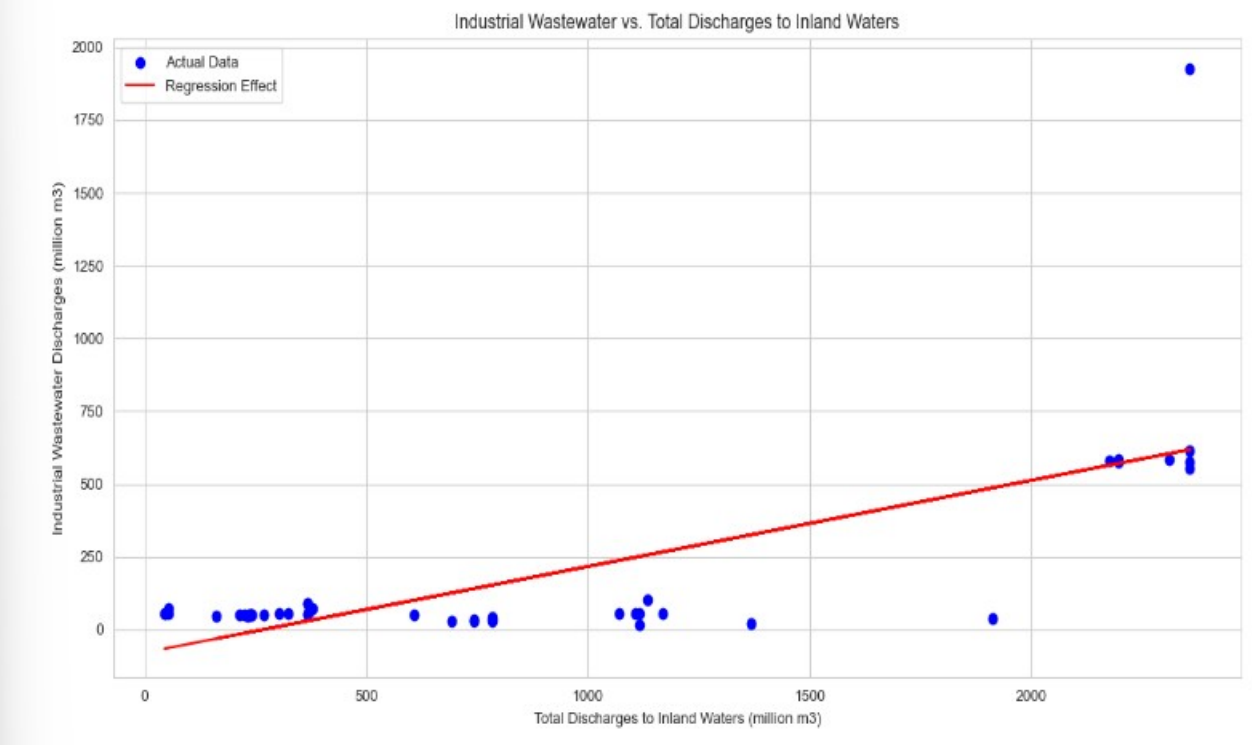


Figure 17

Random Forest Regression

The Random Forest regression model was used to predict deaths due to unsafe WASH practices, based on various environmental and infrastructural predictors. This evaluation typically includes metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) to gauge how well the model predicts continuous numerical outcomes. Here are the key outcomes were calculated:

- Mean Squared Error (MSE): The model achieved an MSE of approximately 0.014. While the MSE is lower than what we observed with the linear regression model, it still represents the average of the squares of the errors—the difference between the observed values and the values predicted by the model.

- R-squared (R^2): The R^2 value is approximately -10.72. Like the linear regression model, a negative R^2 value indicates that the model does not fit the data well and performs worse than a simple mean-based model.

The visualization (figure 18) of feature importance from the Random Forest model reveals which predictors have the most influence on predicting deaths due to unsafe WASH practices. The importance values help identify which environmental and infrastructural features contribute most to the model's predictions.

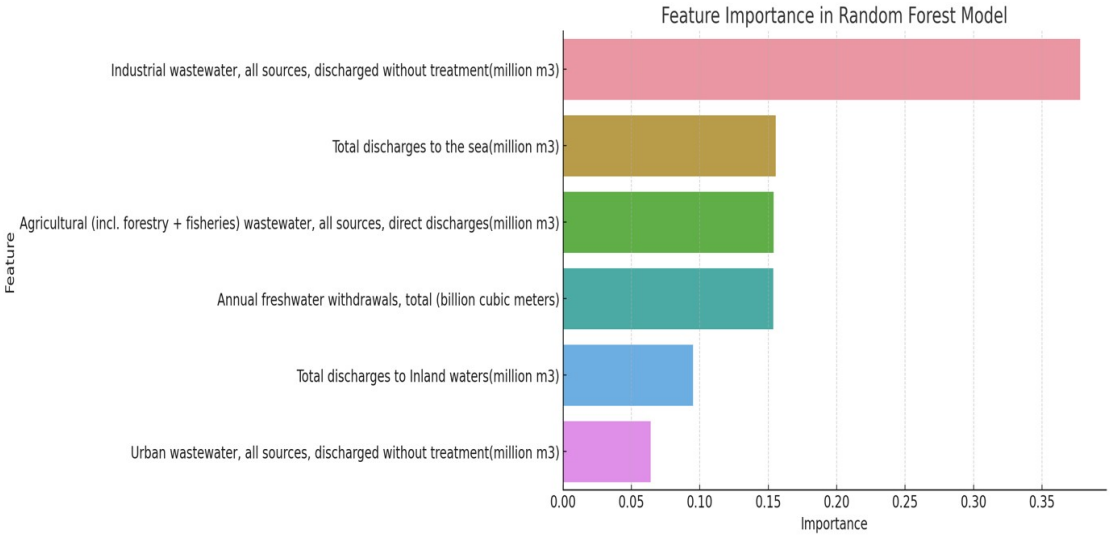


Figure 18

However, the negative R^2 value suggests that the Random Forest model, despite its ability to capture non-linear relationships and interactions between predictors, still does not perform well on this dataset. This could be due to underlying complexities in the data that are not captured by the current set of predictors or due to overfitting to the training data. Moreover, the model's poor performance suggests caution in interpreting these importance values as definitive indicators of causal relationships.

Therefore, linear and Random Forest regression analyses struggled to accurately predict deaths due to unsafe WASH practices based on the selected predictors. This highlights the complexity of the relationship between WASH practices and health outcomes, which may be influenced by a wide range of socioeconomic, environmental, and behavioural factors not fully captured in the current dataset (figure 19).

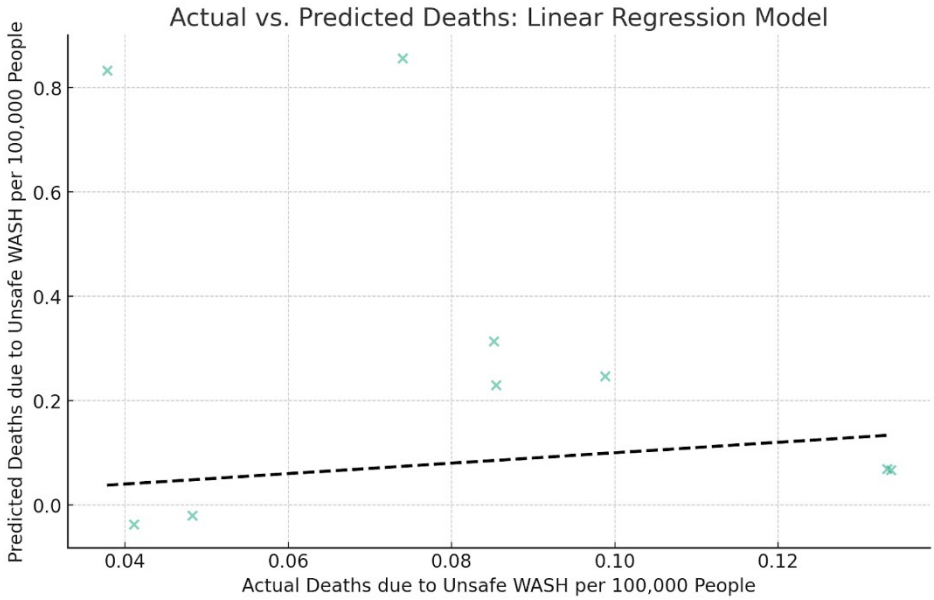


Figure 19

The scarcity of data about the societal impact of wastewater leads the project to look for literature researches and conducted investigations. Through an extensive literature review, evidence is presented to underscore the significance of understanding this impact, laying the groundwork for subsequent analysis and exploration in the research.

High social impact:

The study found that sewage had significant adverse social impacts on the local rural population in Al-Hair, Saudi Arabia. The authors state that "the vast majority of respondents indicated the high impact of sewage on social and economic aspects with the percentages of 85.6% and 84.4% respectively." This overwhelmingly high percentage highlights the substantial negative consequences that inadequate sanitation infrastructure can have on the social well-being of a community (figure 20).

Migration and criminal activity:

Two specific social problems mentioned in the study were that "sewage contributes to immigration from Al-Hair" and "trees around sewage became a place for criminals." The former suggests that the poor sanitation conditions were making the area an undesirable place to live, leading to locals migrating away, while the latter indicates that the sewage was creating an environment conducive to illegal activities.

Demographic factors influencing perceptions:

Interestingly, the study found that perceptions of these impacts varied based on demographic factors. The authors state that "the interrelation between the perception of the diverse effects of sewage and people's characteristics indicate that age, gender, household size and education level, are key determinants of rural people' perception on health, social and economic-related risks due to sewage." Older respondents (over 50 years), females, those with university degrees, and those with larger families (>8 members) tended to have a higher awareness of the adverse social effects of sewage. This suggests that these groups may be more attuned to the social consequences of poor sanitation, possibly due to factors such as greater concern for family well-being, higher education levels, and more life experience.

Need for targeted policies and education:

Given these varying perceptions among different demographic groups, the authors recommend that "relevant policies are required to minimize the different hazards of sewage, keeping in view the socio-economic characteristics of the populations living near the sewage facilities." They also emphasize the need for educational initiatives, stating that "the study also establishes the need for the launching of the Extension and Education programs to create awareness on the adverse effects of sewage and strategies to reduce their harmful effects." Such targeted policies and awareness-raising programs could help to mobilize community support for addressing the issue.

Importance of stakeholder involvement:

Finally, the authors recommend a participatory approach to managing sewage, noting that "additional research is required to suggest intervention framework for dealing with sewage by involving all stakeholders in the management of sewage to ensure sustainable development." This stakeholder involvement could help to ensure that the needs and concerns of the local community are considered, and that any solutions are socially and economically sustainable in the long term (Aldosari, et al,. 2017)

In summary, the study provides clear evidence of the significant negative social impacts of sewage on the rural community in Al-Hair, with perceptions of these impacts varying based on demographic factors. The authors' recommendations for targeted policies, educational programs, and a multi-stakeholder management approach offer potential strategies for addressing this complex issue and promoting sustainable development and social well-being.

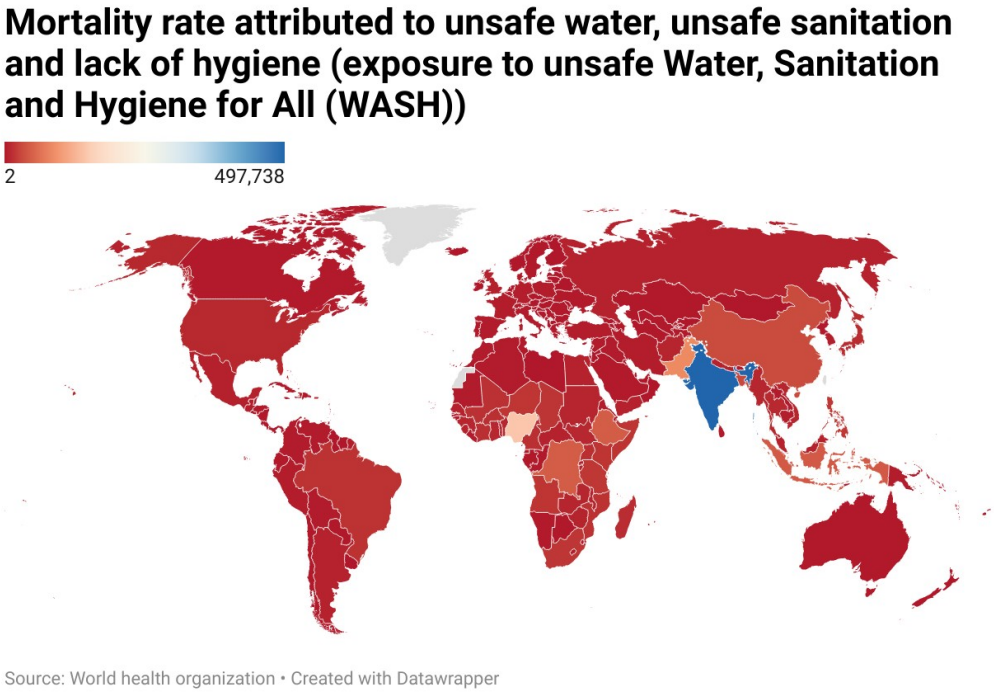


Figure 20

In order to find wastewater impact on carbon density, the same approach was applied as on impact on society.

According to published paper in 2022, “Carbon neutrality of wastewater treatment - A systematic concept beyond the plant boundary” achieving carbon neutrality in wastewater treatment extends beyond the treatment plant boundary and requires a systematic approach. As stated in the document, "carbon neutrality of the wastewater system is far beyond the plant boundary".

Figure 21 in the document provides a sketch of multiple boundaries for carbon accounting of wastewater treatment. It illustrates three boundaries: within-the-fence of WWTPs (yellow dashed line), urban infrastructure related to WWTPs (pink dashed line), and human society and ecological system (blue dashed line). The figure shows that the scope of carbon accounting can be gradually expanded from wastewater treatment processes to the entire human society and ecological system. It also depicts direct carbon emissions with specific symbols and indirect carbon emissions implied in the flows of energy/chemicals consumption and recycled products.

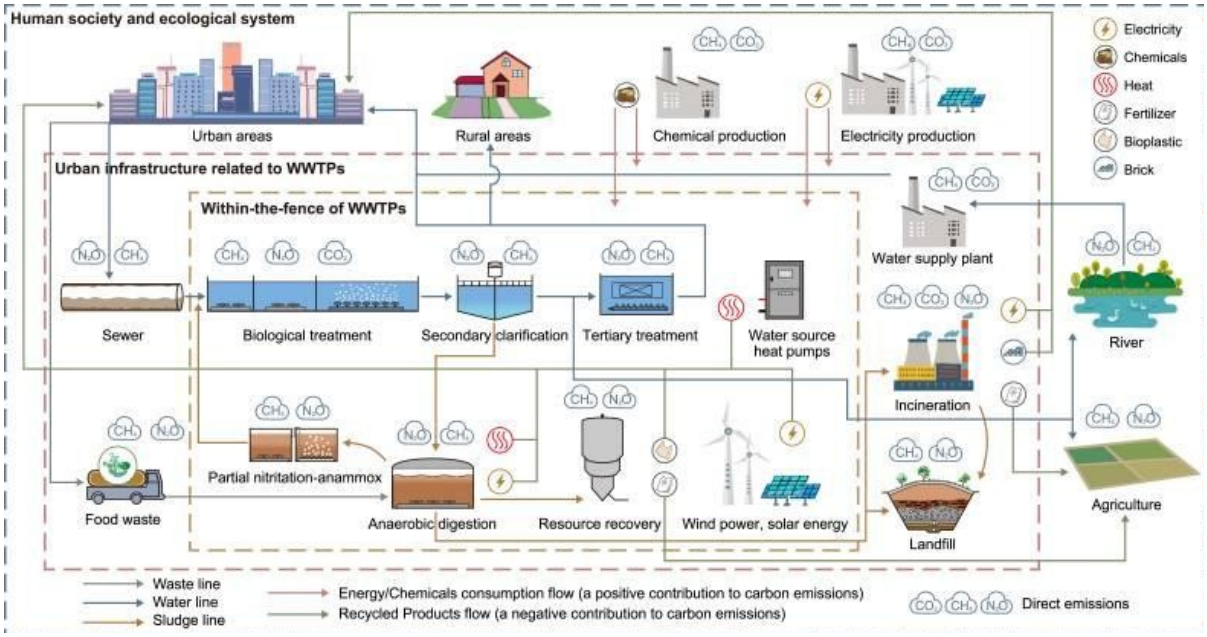


Figure 21

Key points and recommendations from other sources:

1. "Wastewater treatment can contribute approximately 1–2% of the total greenhouse gas (GHG) emissions in the world". To achieve carbon neutrality, a comprehensive accounting of direct and indirect carbon emissions within and beyond wastewater treatment plants (WWTPs) is necessary (Lu, Guest et al., 2018)

2. Direct emissions include CH4 and N2O from wastewater treatment processes and sludge treatment/disposal. The document recommends "including fossil CO2 in GHG accounting is necessary for setting accurate guidelines" (IPCC, 2013).

3. Indirect emissions stem from energy and resource consumption. Energy-saving measures, such as "upgrading obsolete equipment and apply real-time controllers" (McCarty, Bae, 2011) can reduce indirect emissions.

4. Energy recovery through biogas production from anaerobic digestion and water source heat pumps can offset energy demand. However, the document notes that "energy neutrality and carbon neutrality are two different terms" and "successful wastewater treatment cases in energy self-sufficiency may not achieve carbon neutrality" (McCarty, Bae, 2011).

5. Resource recovery, specially producing carbon-based materials like bioplastics and biochar, is encouraged. The document states that "the production of carbon-based materials from wastewater treatment can promote CO2 sequestration and decrease GHG emissions" (McCarty, Bae, 2011).

6. Decentralized wastewater treatment systems are highlighted as a potential solution for reducing energy consumption, enhancing resource recovery, and closing the water loop. The document mentions that "decentralized system outcompetes the centralized system in less energy input and more efficient resource recovery" (Hao, Liu, 2015).

In summary, achieving carbon neutrality in wastewater treatment requires a holistic approach that considers direct and indirect emissions, energy and resource recovery, decentralized systems, and the overall urban infrastructure layout. Figure 1 illustrates the multiple boundaries for carbon accounting, emphasizing the need to extend beyond the treatment plant boundary. Addressing these challenges is essential for reducing the carbon intensity of the wastewater sector and promoting social well-being through improved sanitation and public health.

Insights and recommendations

The project focused on water contaminants has shed light on the pivotal role of wastewater as the strongest global metric impacting various facets of the environment and human health. It was found that by increasing the volume of urban wastewater premature deaths rise by approximately 0.844, highlighting significant health risks from inadequate wastewater management. Moreover, discharge volumes of urban, industrial and agricultural wastewater have a direct impact on total wastewater discharges to sea or inland waters resulting in water pollution around the globe.

Through rigorous analysis and literature review, it became evident that wastewater exerts a significant influence not only on environmental parameters but also on societal well-being. The findings underscore the interconnectedness between wastewater contamination and adverse impacts on ecosystems, public health, and societal dynamics. Furthermore, insights gleaned from the study indicate a correlation between wastewater pollution and carbon

density, highlighting the broader implications for climate change mitigation efforts. By recognizing wastewater as a critical determinant of environmental sustainability and public health, the project underscores the urgency of addressing wastewater management challenges and implementing effective mitigation strategies.

Recommendations for Koru Impact Solutions:

1. Acquire Recent and Accurate Data:

- Invest in acquiring updated wastewater data beyond 2021 to ensure the accuracy and relevance of analyses. Collaborate with data providers to access real-time or near-real-time data streams for ongoing monitoring and assessment.

2. Enhance Geographic Granularity:

- Expand data integration efforts to incorporate diverse environmental datasets, including climate, topography, land use, and population density data. This comprehensive approach will enable the identification of areas with significant environmental variations and potential water contamination risks globally.

3. Data Integration and Harmonization:

- Develop robust data integration pipelines and ETL (Extract, Transform, Load) processes to effectively ingest, cleanse, transform, and integrate heterogeneous wastewater datasets from various sources. This will facilitate seamless data interoperability and enhance the accuracy and reliability of analyses.

4. Standardize Data Collection Protocols:

- Establish standardized protocols and formats for gathering wastewater data, ensuring consistency and comparability across different sources. Implement evaluation systems to verify the reliability and quality of data obtained from government agencies, water utilities, and industrial sites.

5. Promote Stakeholder Collaboration:

- Foster collaboration with relevant stakeholders, including government agencies, regulatory bodies, research institutions, and industry partners, to enhance data collection, sharing, and validation efforts. Leverage partnerships to access additional datasets and expertise for comprehensive analysis.

6. Invest in Data Quality Assurance:

- Allocate resources towards implementing robust data quality assurance mechanisms, including validation processes and quality control measures. Prioritize data accuracy, completeness, and reliability to ensure the integrity of analytical results and insights.

7. Explore Advanced Analytical Techniques:

- Explore the use of advanced analytical techniques, such as machine learning algorithms and predictive modelling, to extract actionable insights from wastewater data. Embrace innovative approaches to uncover hidden patterns, trends, and correlations for informed decision-making.

8. Continual Improvement and Adaptation:

- Embrace a culture of continual improvement and adaptation, fostering agility and responsiveness to evolving data requirements and technological advancements. Regularly review and refine data collection, integration, and analysis processes to enhance effectiveness and efficiency over time.

By implementing these recommendations, Koru Impact Solutions can enhance its data-driven approach to sustainability impact assessment, enabling informed decision-making and promoting positive environmental outcomes for its stakeholders and partners.

Given more time and project extension, it's worth noting the potential for implementing additional statistical analyses to enhance the accuracy and depth of insights. Specifically, techniques such as non-linear regression and residual vs. fitted model analysis could be explored to refine the analytical process further. These advanced methods offer opportunities to capture complex relationships and patterns within the data, potentially yielding more nuanced and precise results. By incorporating such approaches, the project can elevate its analytical rigour and contribute to a more comprehensive understanding of the factors influencing water contamination and its impacts.

Github link: https://github.com/YahyaHabib/-Water---contaminants-and-levels

Individual reflections

Deadline set up for each team member for the individual reports was 29/03/24 20:00.

Laurita Kunickaite

As a team leader, my contributions have played a pivotal role in steering our project towards success. A concise summary of my key roles and achievements entails meticulous project planning, proficient facilitation of team meetings, adept client communication, and insightful recommendations for business executives. Through the implementation of efficient project management tools and methodologies, I ensured that our project plan comprehensively outlined all tasks, and timelines, thereby highlighting clarity and direction within the team. I provided regular reviews and updates to the ensuring its relevance and effectiveness throughout the project lifecycle.

In addition to project planning, I took charge of organizing and facilitating team meetings, which served as crucial platforms for aligning our collective efforts towards project objectives. By providing structured agendas and meticulously documenting meeting minutes, I ensured that discussions remained focused on key action items derived from our project plan. These meetings not only facilitated effective communication but also fostered a collaborative environment where team members could express their insights and concerns.

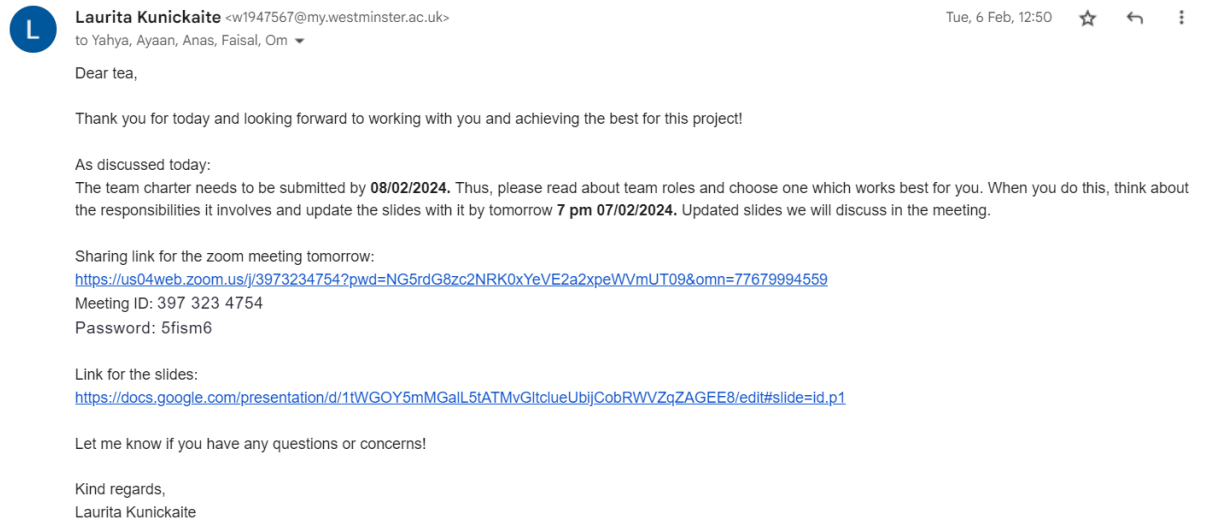
Moreover, maintaining transparent communication with the client was another integral aspect of my role. By providing regular progress updates and addressing client inquiries and concerns promptly, I ensured that their expectations were managed effectively. This open line of communication fostered trust and confidence, laying the groundwork for a successful client-team relationship.

Nevertheless, my journey was not without its challenges. The main obstacle was the initial uncertainty surrounding the client's expectations. Ambiguity regarding project deliverables and objectives posed a significant challenge, requiring proactive measures to overcome. By my established regular communication channels and actively seeking feedback, I was able to clarify expectations and align with my team efforts accordingly.

Furthermore, another challenge for me was the relentless time pressure and workload demands, exacerbated by a lack of expertise in certain areas. Overcoming this challenge I developed a multifaceted approach, including prioritizing tasks, upskilling through collaborative learning, and seeking guidance from module leader and client. As a result, my team not only met deadlines but also delivered outcomes that exceeded expectations.

Despite the challenges, this project served as a profound learning experience for me, enriching my leadership skills and enhancing my ability to navigate uncertainty and guide diverse teams towards success. By trusting my instincts, embracing feedback, and fostering a collaborative environment, I gained invaluable insights that will undoubtedly shape my future endeavors. Also, the positive feedback received from my peers further bolstered my confidence as a leader, reaffirming my commitment to fostering a culture of excellence and collaboration.

Overall, my contributions as a team leader have been crucial in maintaining project clarity, maintaining effective communication, and delivering value to both the client and the organization.



Laurita Kunickaite <w1947567@my.westminster.ac.uk> Tue, 6 Feb, 12:50
to Yahya, Ayaan, Anas, Faisal, Om

Dear tea,

Thank you for today and looking forward to working with you and achieving the best for this project!

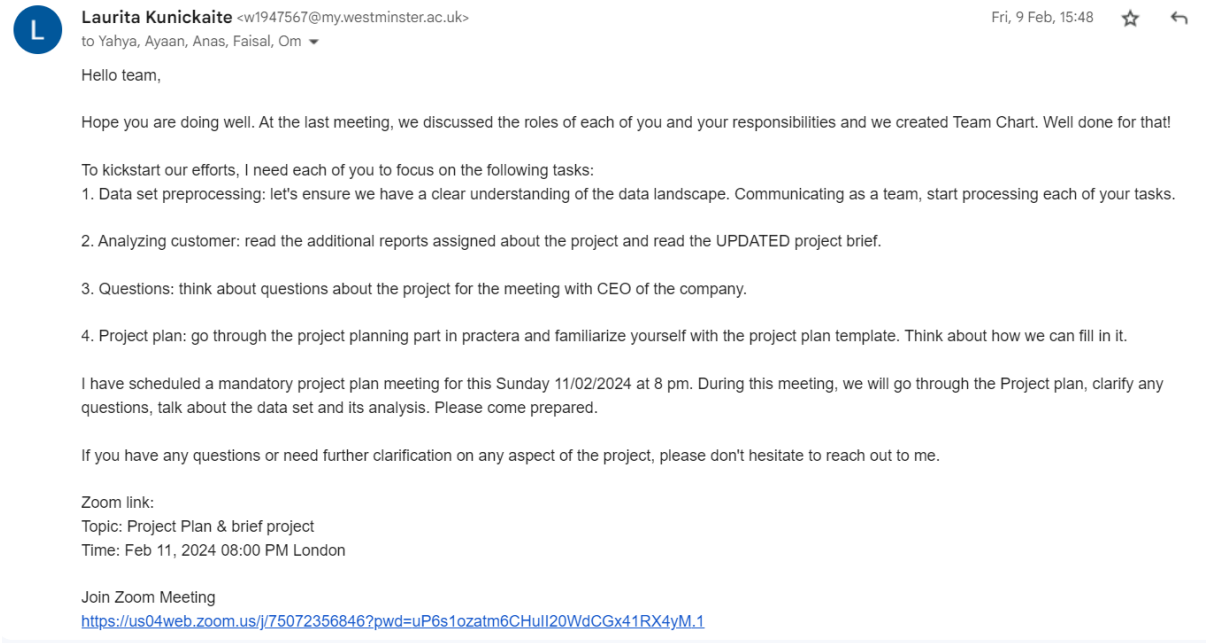
As discussed today:
The team charter needs to be submitted by **08/02/2024.** Thus, please read about team roles and choose one which works best for you. When you do this, think about the responsibilities it involves and update the slides with it by tomorrow **7 pm 07/02/2024.** Updated slides we will discuss in the meeting.

Sharing link for the zoom meeting tomorrow:
https://us04web.zoom.us/j/3973234754?pwd=NG5rdG8zc2NRK0xYeVE2a2xpeWVmUT09&omn=77679994559
Meeting ID: 397 323 4754
Password: 5fism6

Link for the slides:
https://docs.google.com/presentation/d/1tWGOY5mMGalL5tATMvGltclueUbijCobRWVZqZAGEE8/edit#slide=id.p1

Let me know if you have any questions or concerns!

Kind regards,
Laurita Kunickaite



Laurita Kunickaite <w1947567@my.westminster.ac.uk> Fri, 9 Feb, 15:48
to Yahya, Ayaan, Anas, Faisal, Om

Hello team,

Hope you are doing well. At the last meeting, we discussed the roles of each of you and your responsibilities and we created Team Chart. Well done for that!

To kickstart our efforts, I need each of you to focus on the following tasks:
1. Data set preprocessing: let's ensure we have a clear understanding of the data landscape. Communicating as a team, start processing each of your tasks.

2. Analyzing customer: read the additional reports assigned about the project and read the UPDATED project brief.

3. Questions: think about questions about the project for the meeting with CEO of the company.

4. Project plan: go through the project planning part in practera and familiarize yourself with the project plan template. Think about how we can fill in it.

I have scheduled a mandatory project plan meeting for this Sunday 11/02/2024 at 8 pm. During this meeting, we will go through the Project plan, clarify any questions, talk about the data set and its analysis. Please come prepared.

If you have any questions or need further clarification on any aspect of the project, please don't hesitate to reach out to me.

Zoom link:
Topic: Project Plan & brief project
Time: Feb 11, 2024 08:00 PM London

Join Zoom Meeting
https://us04web.zoom.us/j/75072356846?pwd=uP6s1ozatm6CHuII20WdCGx41RX4yM.1

Faisal Qaderi (submitted individual part 29/03/24 at 05:07):

When Yahya and I began our project, we had a big task ahead with six sets of data to sort out. We thought splitting them up would be smart, so each of us took three sets. This seemed like a good way to share the work. However, after showing our first efforts to others, we discovered we were not on the right path. This surprise was helpful because it alerted us early.

Understanding we needed a change, we decided to work together on all the datasets. This approach was far better. Yahya was excellent at leading the clean-up and preparation of the data. I was gearing up for the next phase: analysing the cleaned data to make sense of it through statistics.

Once Yahya, om and I had the data in good shape, i then went on to take over, diving deep into the numbers to identify patterns. I explored various analysis methods and compiled a report on my findings. However, we then received more feedback, highlighting areas for improvement. With Yahya's support, I was able to refine my work.

During this project, we faced several challenges. The first was organizing all the data, a massive job that showed us the value of teamwork. The second challenge was the detailed analysis and interpretation of the data. This required a lot of learning and extra effort but was deeply rewarding in the end.

This project was a rich learning journey. It taught me about the subject of the project and offered valuable life lessons in remaining calm under pressure, staying organized, and improving my teamwork skills. It felt like we were performing a real, professional task, which taught me significant lessons. Receiving advice and comments from our colleagues was incredibly valuable. It provided new perspectives and strengthened my contributions.

In total, this project was an important learning experience. Facing challenges and receiving guidance from our team helped a lot. Working closely with Yahya and the rest of the team emphasized the importance of listening and collaborating effectively.

Initially, splitting the datasets seemed logical for efficiency. However, the feedback we received was a wake-up call, showing us that our separate efforts were misaligned. This led to a pivotal shift in our strategy, where collaboration became our new mantra. Merging our efforts not only streamlined the process but also enhanced the quality of our work,

Dealing with feedback constructively was another critical lesson. It pushed us to revisit our work, not as a setback but as an opportunity for improvement. This iterative process of review and refinement was crucial in driving our project forward.

Moreover, the project served as a practical lesson in dealing with complex challenges. From the daunting task of data organization to the intricate details of data analysis, each step was a building block in our learning journey. These experiences underscored the essence of perseverance, critical thinking, and the willingness to learn from feedback.

Collaborating on this project also improved my interpersonal skills. Working in a team setting, especially in facing challenges and integrating feedback, taught me the value of diverse viewpoints. It reinforced that constructive criticism is not just about pointing out flaws but about offering pathways to improvement.

In conclusion, this project was not just about data analysis; it was a comprehensive learning experience that extended beyond technical skills to include personal growth and teamwork. It was a journey that showcased the power of collaboration, the importance of resilience, and the value of embracing feedback for continuous improvement.

Om Sadigale (submitted on 29/03/24 00:02):

During my role, I was tasked with leading the data evaluation process, which was crucial in setting the stage for our project. As part of the data preprocessing team, I played a pivotal role in outlining the selection criteria for datasets, countries, and the necessary steps for data cleaning and evaluation. We encountered several challenges, particularly in deciding which datasets and countries to prioritize due to missing data and diverse sources. Despite tight deadlines, we managed to navigate through these challenges effectively, dedicating ample time to preprocessing and cleaning the data.

Initially, I struggled with communication, feeling nervous about expressing my doubts to my colleagues. However, as time went on, I became more comfortable sharing my ideas, which not only boosted my confidence but also improved my communication skills significantly. Working in such an

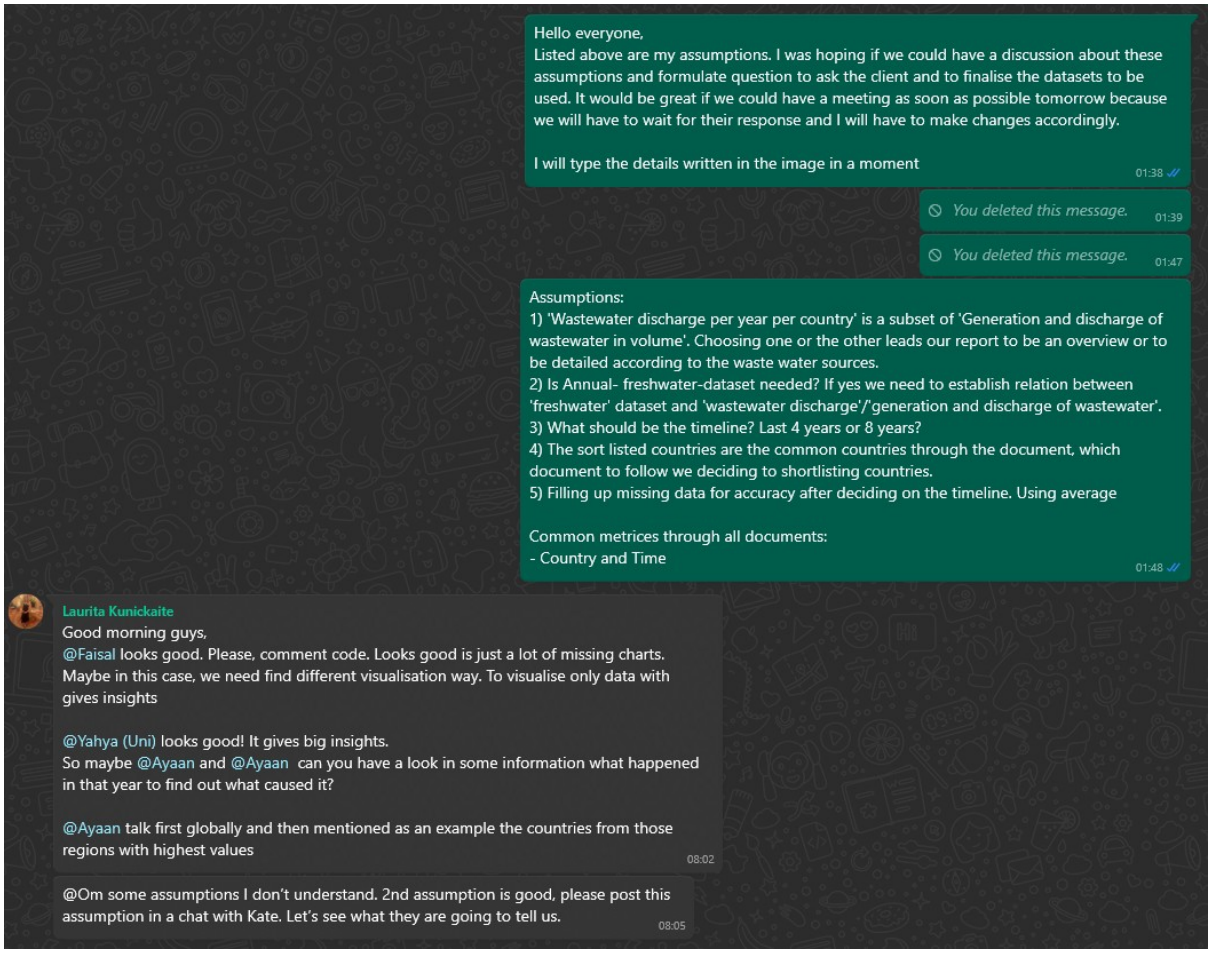
environment taught me invaluable lessons about teamwork and handling pressure, giving me a glimpse into the realities of industry workflows.

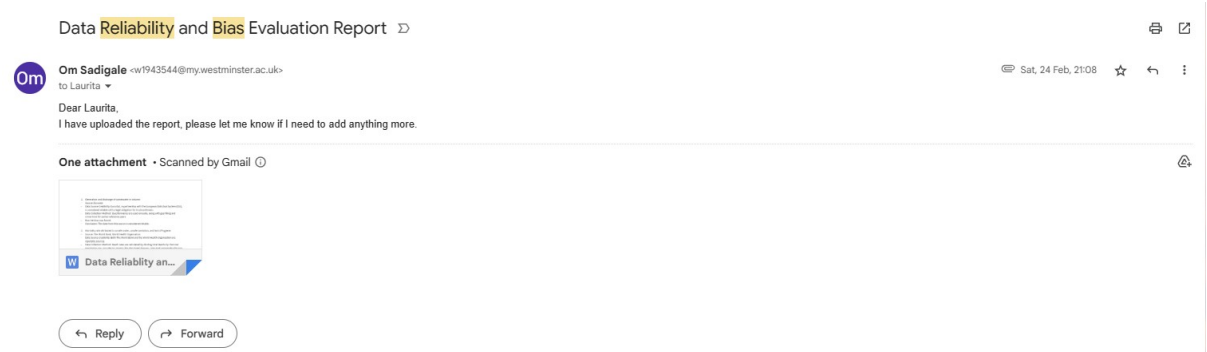
I also had the opportunity to learn from Laurita, who exemplified essential leadership qualities. Her calm demeanor in stressful situations and her ability to strike a balance between kindness and firmness left a lasting impression on me. Rather than panicking, she always focused on finding solutions, which was incredibly inspiring.

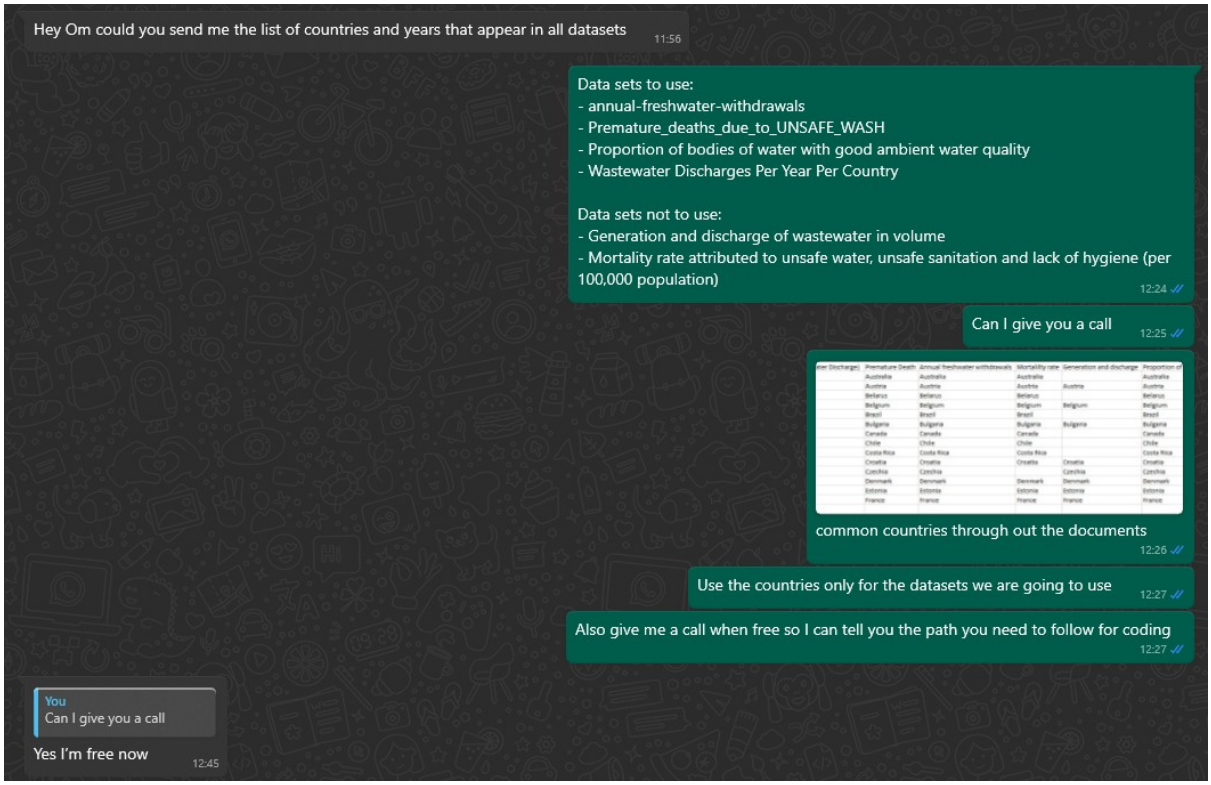
Delivering presentations to clients further honed my communication skills and taught me the importance of clear and concise communication. Additionally, interacting with my teammates on a personal level allowed me to understand them better, ultimately improving our professional coordination and teamwork.

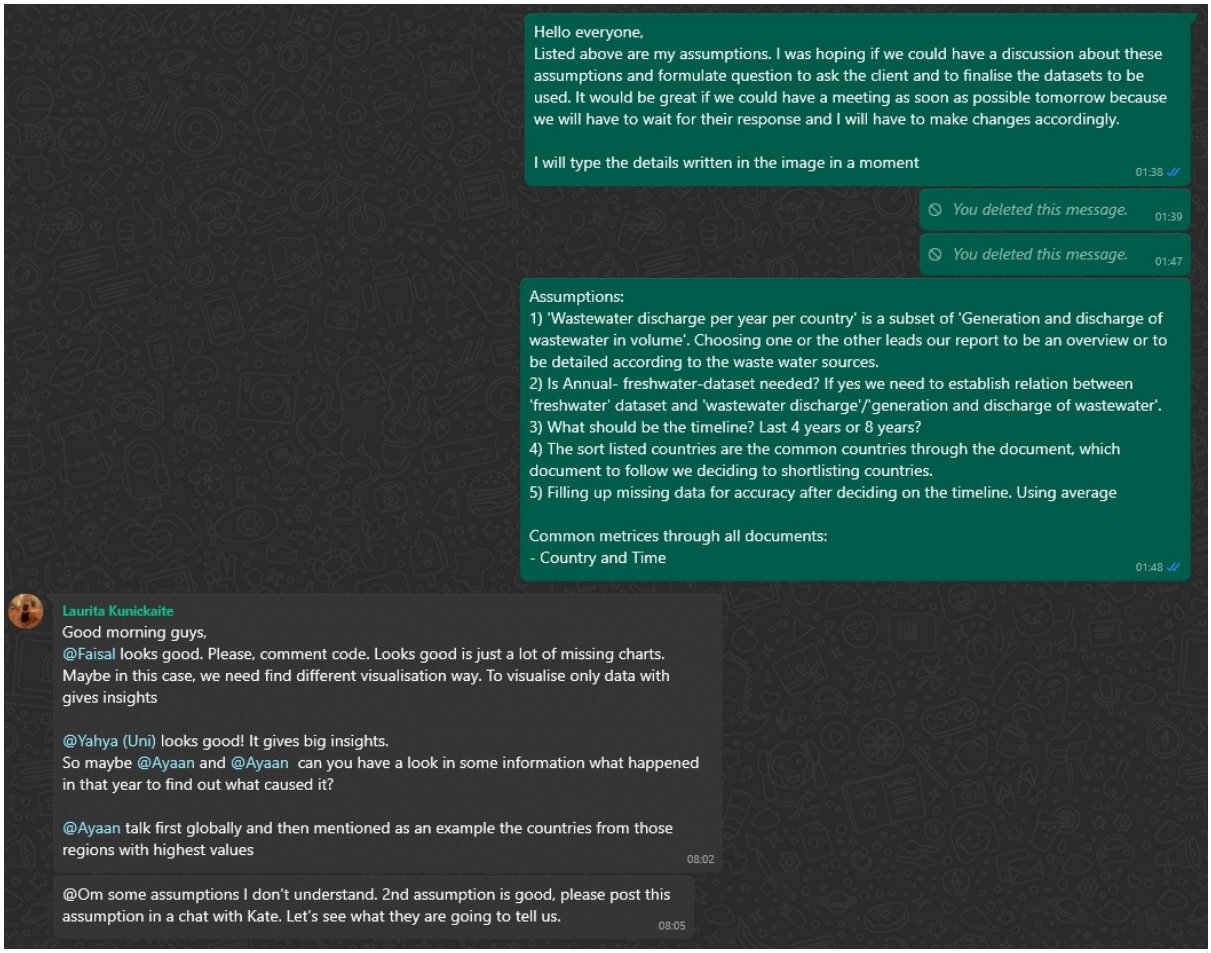
Overall, this experience was transformative, helping me grow both personally and professionally, and providing me with invaluable insights into the dynamics of collaborative work environments.

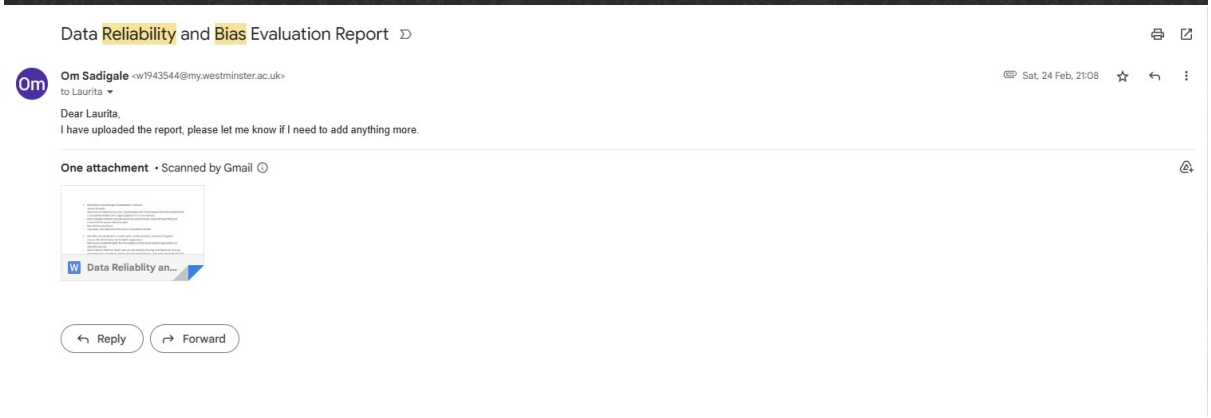
Evidence:

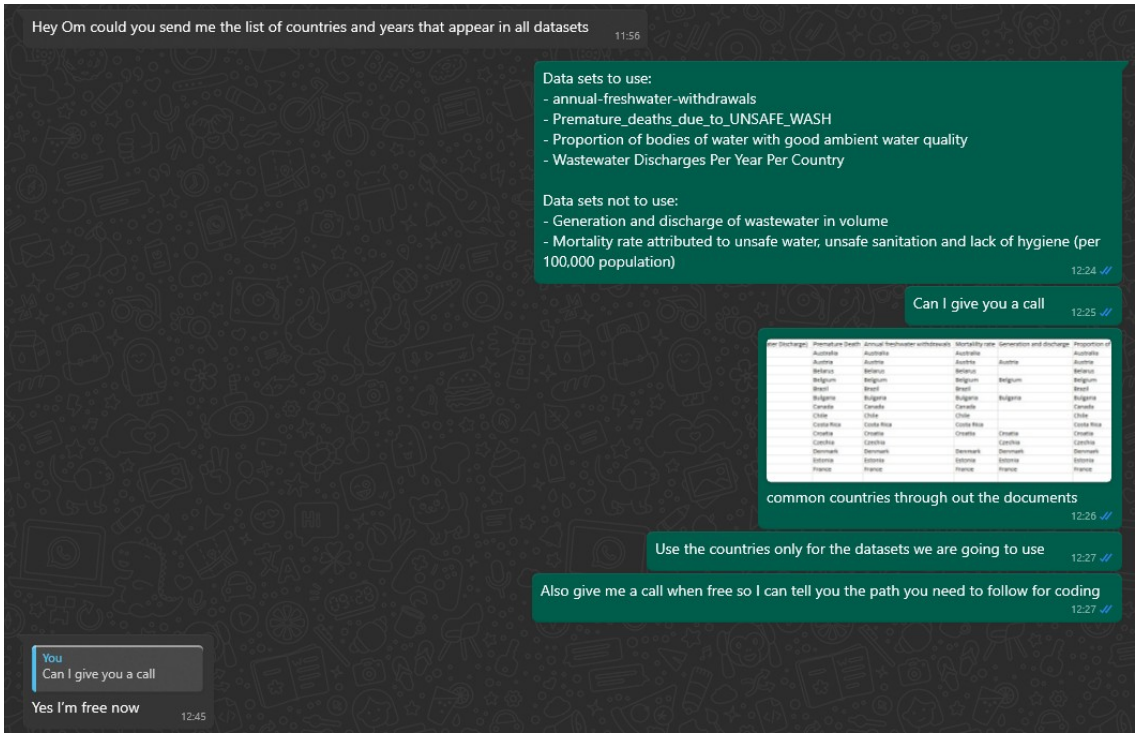



Data Reliability and Bias Evaluation Report
Om Sadigale <w1943544@my.westminster.ac.uk>
to Laurita
Sat, 24 Feb, 21:08
Dear Laurita,
I have uploaded the report, please let me know if I need to add anything more.
One attachment • Scanned by Gmail
Data Reliablity an...
Reply
Forward


Hey Om could you send me the list of countries and years that appear in all datasets
11:56
Data sets to use:
- annual-freshwater-withdrawals
- Premature_deaths_due_to_UNSAFE_WASH
- Proportion of bodies of water with good ambient water quality
- Wastewater Discharges Per Year Per Country
Data sets not to use:
- Generation and discharge of wastewater in volume
- Mortality rate attributed to unsafe water, unsafe sanitation and lack of hygiene (per 100,000 population)
12:24
Can I give you a call
12:25
common countries through out the documents
12:26
Use the countries only for the datasets we are going to use
12:27
Also give me a call when free so I can tell you the path you need to follow for coding
12:27
You
Can I give you a call
Yes I'm free now
12:45


Hello everyone,
Listed above are my assumptions. I was hoping if we could have a discussion about these assumptions and formulate question to ask the client and to finalise the datasets to be used. It would be great if we could have a meeting as soon as possible tomorrow because we will have to wait for their response and I will have to make changes accordingly.
I will type the details written in the image in a moment
01:38
You deleted this message.
01:39
You deleted this message.
01:47
Assumptions:
1) 'Wastewater discharge per year per country' is a subset of 'Generation and discharge of wastewater in volume'. Choosing one or the other leads our report to be an overview or to be detailed according to the waste water sources.
2) Is Annual- freshwater-dataset needed? If yes we need to establish relation between 'freshwater' dataset and 'wastewater discharge'/'generation and discharge of wastewater'.
3) What should be the timeline? Last 4 years or 8 years?
4) The sort listed countries are the common countries through the document, which document to follow we deciding to shortlisting countries.
5) Filling up missing data for accuracy after deciding on the timeline. Using average
Common metrices through all documents:
- Country and Time
01:48
Laurita Kunickaite
Good morning guys,
@Faisal looks good. Please, comment code. Looks good is just a lot of missing charts. Maybe in this case, we need find different visualisation way. To visualise only data with gives insights
@Yahya (Uni) looks good! It gives big insights.
So maybe @Ayaan and @Ayaan can you have a look in some information what happened in that year to find out what caused it?
@Ayaan talk first globally and then mentioned as an example the countries from those regions with highest values
08:02
@Om some assumptions I don't understand. 2nd assumption is good, please post this assumption in a chat with Kate. Let's see what they are going to tell us.
08:05


Data Reliability and Bias Evaluation Report
Om Sadigale <w1943544@my.westminster.ac.uk>
to Laurita
Sat, 24 Feb, 21:08
Dear Laurita,
I have uploaded the report, please let me know if I need to add anything more.
One attachment • Scanned by Gmail
Data Reliablity an...
Reply
Forward



Yahya Habib (submitted on 29/03/24 16:16):

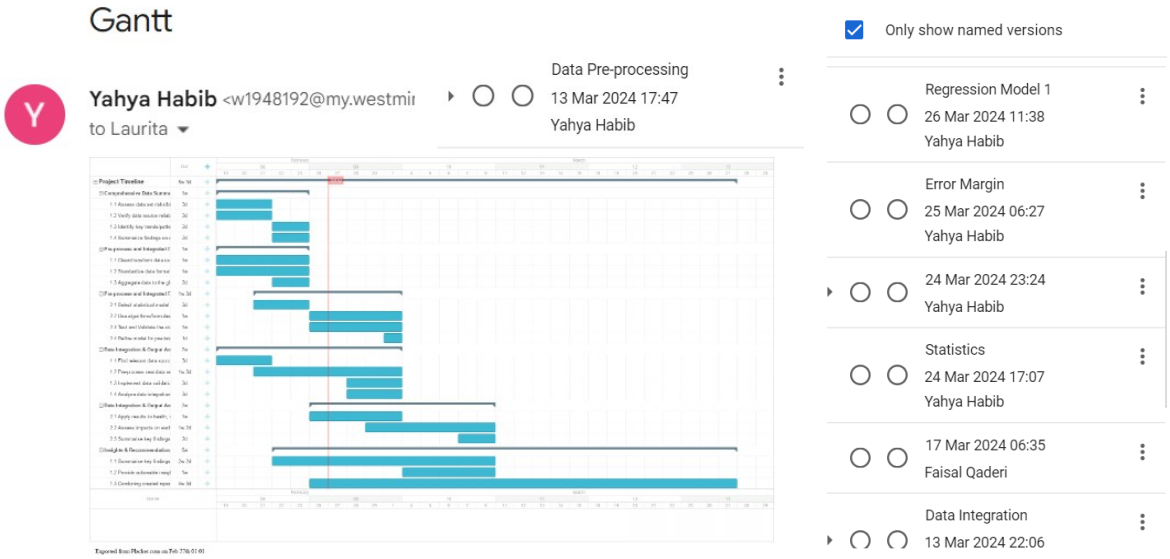
My role shifted from statistical modelling from the original plan to leading the pre-processing phase, aligning with feedback for better data preparation. Despite the change, my contributions remained focused on ensuring data quality and integration. While my tasks differed, the technical implementation using Python stayed consistent with the original plan. I always communicated with team members to ensure we are on all on the same page.

To prepare the data for analysis, I:

- Developed a pre-processing plan to ensure a structured approach.
- Presented the raw data for initial assessment.
- Identified the most relevant variables using insights from existing literature reviews. I consulted Faisal and Om on this section in order to have the best approach on our project.
- Pre-processed the relevant datasets, addressing missing values, duplicates, and errors.
- Integrated the datasets while maintaining the integrity of key variables.

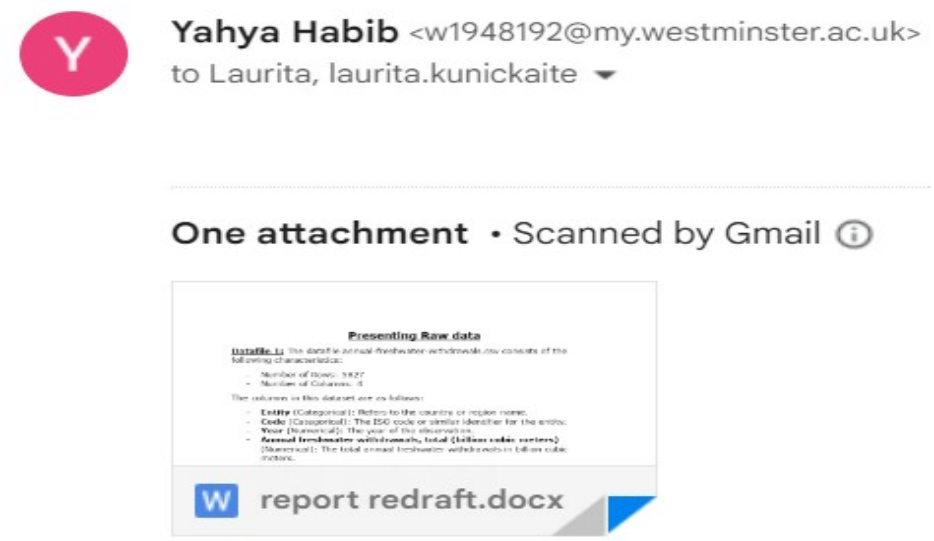
Technical Process Using Python:

- Created heatmaps to visualize data availability for the relevant datasets.
- Cleaned datasets by replacing missing values, handling duplicates.
- Integrated datasets by merging them based on common variables like country and year.
- Conducted further analysis, such as calculating statistics like mean and margin of error for different metrics and performing correlation and regression analyses on Health-related variables in order to get a global metric for health impact.
- Assessed model performance using metrics like the root mean squared error (RMSE) to evaluate predictive accuracy.

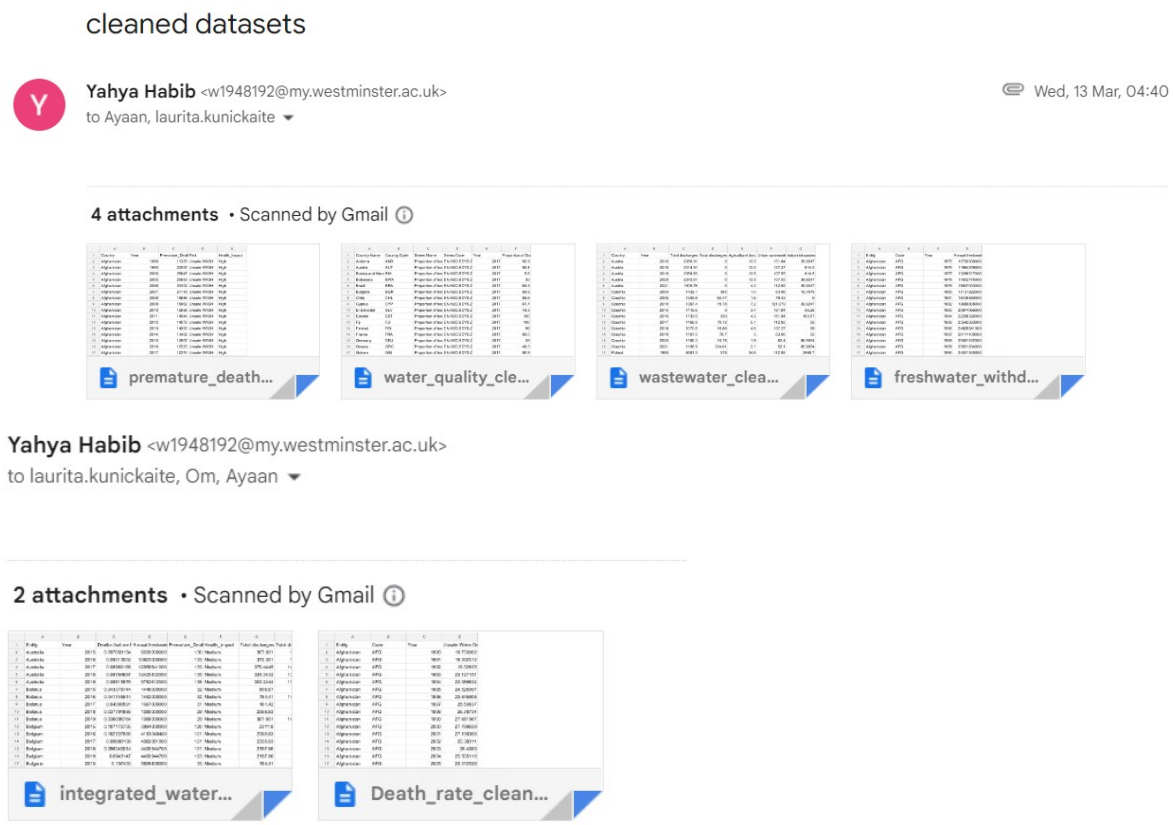


Email containing: Gantt Chart Created

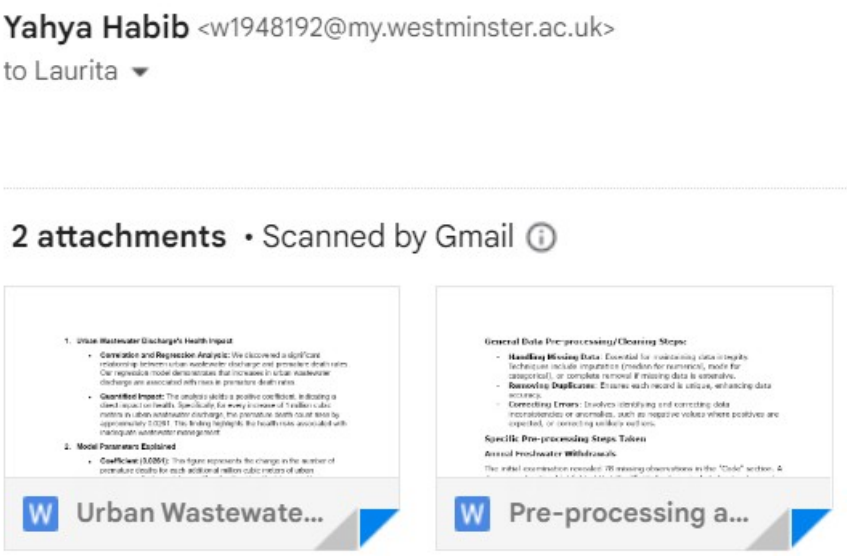
Email containing: File containing the steps taken in preprocessing and integrating data



Email containing: Files containing the cleaned datasets and integrated datasets.



Email containing: Final preprocess draft and the statistical model of Urban Wastewater vs premature death. Which allowed us to measure the health impact



One challenge we faced was the initial data format, which was wide and contained numerous missing values. To address this, we transformed the data to a long format, enabling better analysis. Additionally, inconsistencies in data labelling across datasets posed another problem. However, through vigorous data cleaning and careful handling, we successfully managed to ensure accuracy in our analysis. Another challenge involved ensuring alignment among team members regarding task understanding and project goals. To mitigate this, we held frequent meetings to facilitate clear communication and ensure everyone understood the task and deliverables consistently. These discussions allowed a shared understanding and helped us stay on the same page throughout the project.

Through my work-based learning experience, I've gained a deeper perspective into the extensive process of data cleaning. realising its significance in ensuring accurate analysis. Moreover, I've improved problem-solving skills navigating challenges like data inconsistencies. Effective communication and teamwork were vital, as regular meetings ensured alignment among team members. Project management skills were also developed, balancing tasks and deadlines efficiently. Overall, this experience has deepened my understanding of data analysis while fostering essential professional skills crucial for success in any collaborative environment.

Through teamwork, I learnt how to convey difficult data ideas to peers with various technical backgrounds, allowing me to better explain my approach and include different viewpoints. I also grew better at allocating responsibilities based on my capabilities and negotiating solutions when opposing viewpoints surfaced. Furthermore, peer feedback was important. It helped me detect problems, find alternate solutions, and eventually build my confidence in my data analytic abilities. Overall, I learned the value of clear communication, open-mindedness, and constructive feedback while developing collaborative abilities for future success.

Anas Kigigi (submitted on 29/03/24 20:01):

Introduction

In the dynamic and complex project of analysing water contaminants and their levels for Koru Impact Solutions, I embarked on the journey as a Data Researcher. My primary responsibility was to unearth additional data resources to enrich our dataset, thereby enhancing the accuracy and comprehensiveness of our analysis. This reflection delves into my contributions, the challenges encountered, and the profound learning experiences gained throughout this process.

Contributions

My initial task involved a meticulous evaluation and selection of data sources that could complement our primary dataset provided by Koru Impact Solutions. Recognizing the criticality of data integrity, I leveraged resources like OECD. Stat, Eurostat, and World Health Organization databases to gather credible and relevant data on water pollution and its socio-economic impacts. This endeavour not only broadened our analytical scope but also fortified the reliability of our findings, as mentioned in the project report.

In collaboration with the team, I applied a systematic approach to data pre-processing, ensuring that the integrated datasets were devoid of inconsistencies and ready for analysis. This meticulous preparation was pivotal in facilitating seamless data analysis and modelling phases, underscoring the importance of foundational data work in the success of any data science project.

Challenges

The journey was not devoid of hurdles. One significant challenge was the integration of diverse datasets, each with its unique structure and quality issues. The complexity of merging data from various sources to form a coherent and analysable set required innovative problem-solving and adaptability. Additionally, navigating the vast landscape of public data repositories to find relevant and reliable data necessitated a discerning eye and critical evaluation skills, ensuring that only the most pertinent data were included in our analysis.

Another challenge stemmed from the collaborative nature of the project. Coordinating with team members across different roles, each with their unique perspectives and expertise, demanded effective communication and teamwork. It was essential to align our individual contributions towards a unified goal, a process that, while challenging, was highly rewarding.

Learnings

This project served as a fertile ground for both professional and personal growth. Academically, it reinforced the critical importance of data quality and integrity in the data science lifecycle. I learned advanced techniques in data pre-processing and integration, skills that are indispensable in the field
of data science. Professionally, the project honed my ability to communicate complex data concepts succinctly, an invaluable skill when collaborating with stakeholders and team members.

The project also imparted the importance of resilience and adaptability in problem-solving. Faced with unforeseen challenges, the ability to pivot and find alternative solutions was crucial. This experience has equipped me with a more agile and solution-oriented mindset, valuable in any professional setting.

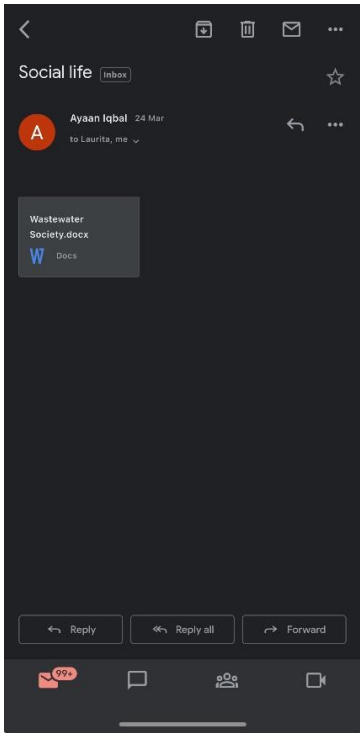
Conclusion

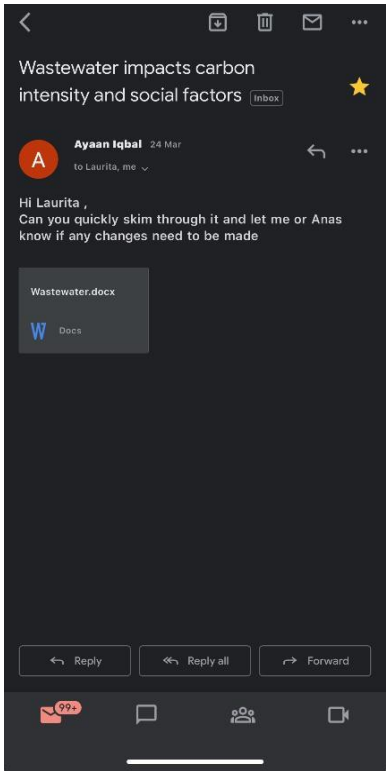
Reflecting on my journey as a Data Researcher within this project, I am immensely grateful for the learning opportunities it presented. The challenges faced were not merely obstacles but stepping stones that enriched my academic and professional journey. The collaborative experience enhanced my teamwork and communication skills, preparing me for future endeavours in the data science domain.

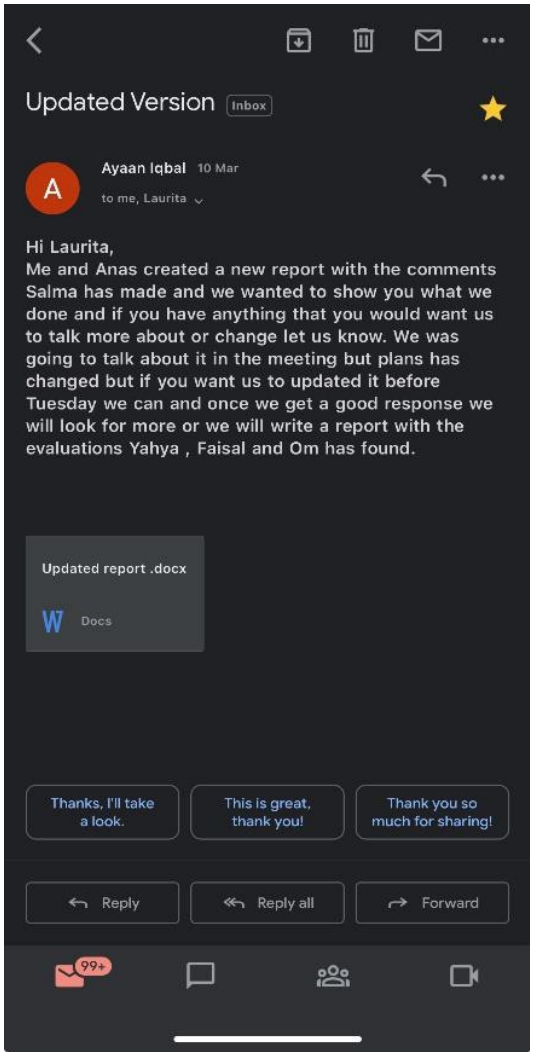
As I move forward, the insights gained from this project will serve as a guidepost, reminding me of the power of data when wielded with integrity and the profound impact of collaborative effort in achieving common goals. This project was not just an academic requirement; it was a transformative experience that has left an indelible mark on my professional development journey.

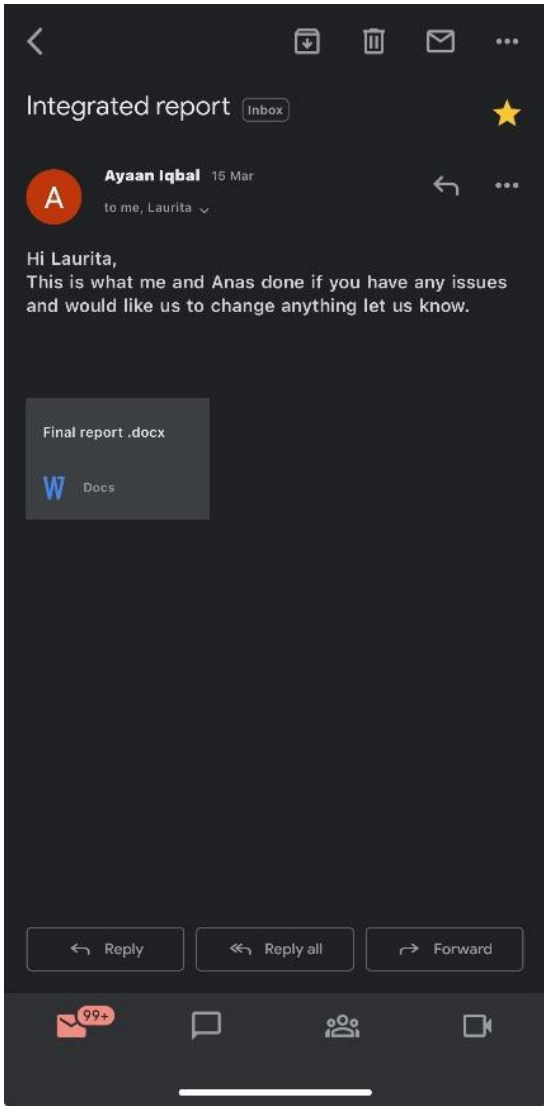
Some proofs of working with my colleague Ayaan on the project :

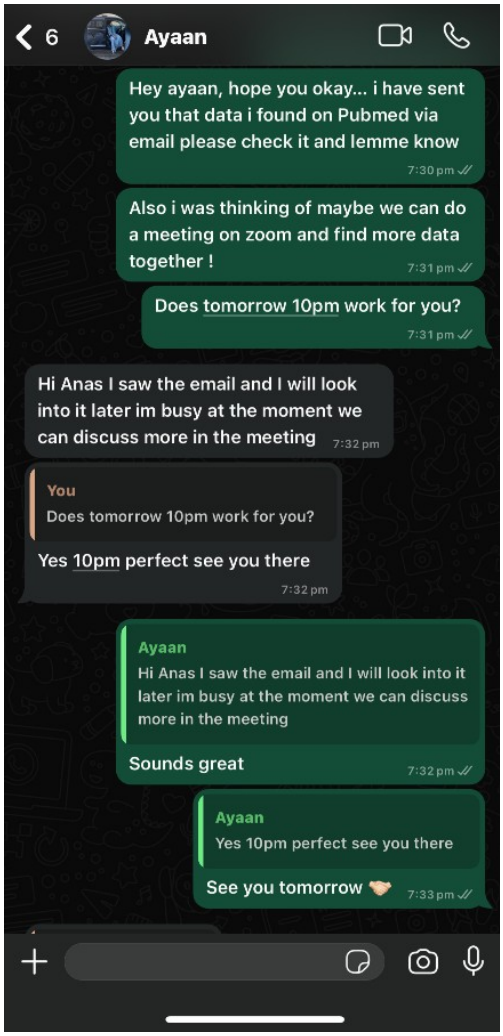
Emails :

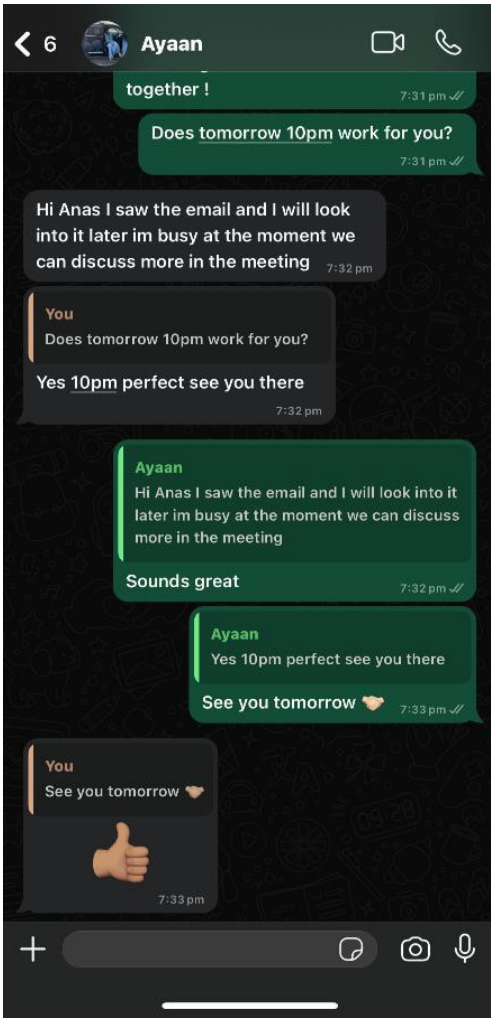

Social life
Inbox
Ayaan Iqbal 24 Mar
to Laurita, me
Wastewater Society.docx
Docs
Reply
Reply all
Forward

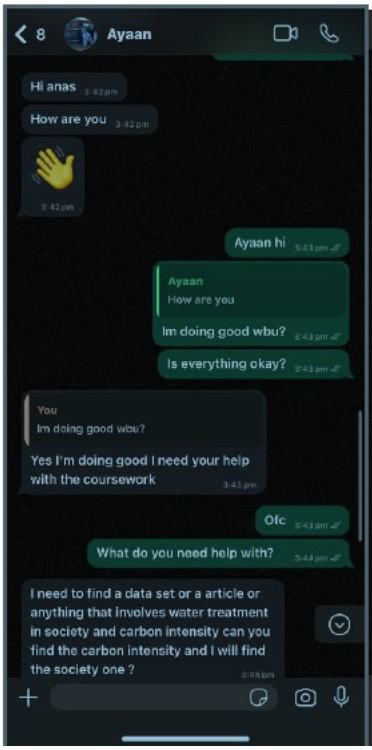

Wastewater impacts carbon intensity and social factors
Inbox
Ayaan Iqbal 24 Mar
to Laurita, me
Hi Laurita ,
Can you quickly skim through it and let me or Anas know if any changes need to be made
Wastewater.docx
Docs
Reply
Reply all
Forward

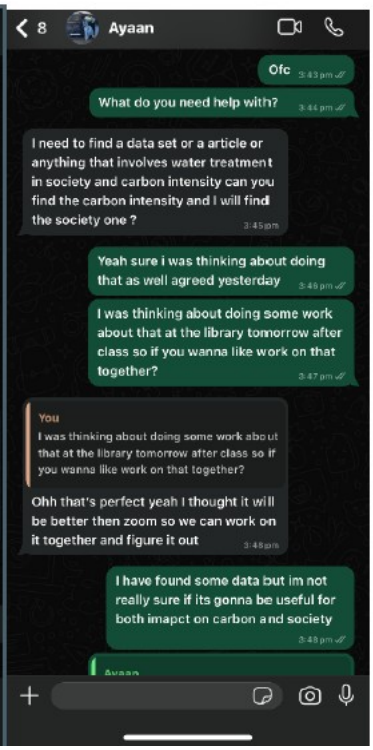

Updated Version
Inbox
Ayaan Iqbal 10 Mar
to me, Laurita
Hi Laurita,
Me and Anas created a new report with the comments Salma has made and we wanted to show you what we done and if you have anything that you would want us to talk more about or change let us know. We was going to talk about it in the meeting but plans has changed but if you want us to updated it before Tuesday we can and once we get a good response we will look for more or we will write a report with the evaluations Yahya , Faisal and Om has found.
Updated report .docx
Docs
Thanks, I'll take a look.
This is great, thank you!
Thank you so much for sharing!
Reply
Reply all
Forward

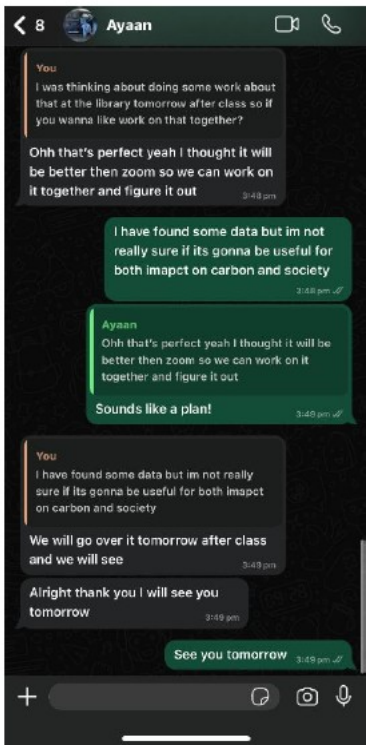

Integrated report
Inbox
Ayaan Iqbal 15 Mar
to me, Laurita
Hi Laurita,
This is what me and Anas done if you have any issues and would like us to change anything let us know.
Final report .docx
Docs
Reply
Reply all
Forward


6
Ayaan
Hey ayaan, hope you okay... i have sent you that data i found on Pubmed via email please check it and lemme know
7:30 pm
Also i was thinking of maybe we can do a meeting on zoom and find more data together !
7:31 pm
Does tomorrow 10pm work for you?
7:31 pm
Hi Anas I saw the email and I will look into it later im busy at the moment we can discuss more in the meeting
7:32 pm
You
Does tomorrow 10pm work for you?
Yes 10pm perfect see you there
7:32 pm
Ayaan
Hi Anas I saw the email and I will look into it later im busy at the moment we can discuss more in the meeting
Sounds great
7:32 pm
Ayaan
Yes 10pm perfect see you there
See you tomorrow
7:33 pm


6
Ayaan
together !
7:31 pm
Does tomorrow 10pm work for you?
7:31 pm
Hi Anas I saw the email and I will look into it later im busy at the moment we can discuss more in the meeting
7:32 pm
You
Does tomorrow 10pm work for you?
Yes 10pm perfect see you there
7:32 pm
Ayaan
Hi Anas I saw the email and I will look into it later im busy at the moment we can discuss more in the meeting
Sounds great
7:32 pm
Ayaan
Yes 10pm perfect see you there
See you tomorrow
7:33 pm
You
See you tomorrow
7:33 pm


8
Ayaan
Hi anas
How are you
Ayaan hi
Ayaan
How are you
Im doing good wbu?
Is everything okay?
You
Im doing good wbu?
Yes I'm doing good I need your help with the coursework
Ofc
What do you need help with?
I need to find a data set or a article or anything that involves water treatment in society and carbon intensity can you find the carbon intensity and I will find the society one ?


8
Ayaan
Ofc
What do you need help with?
I need to find a data set or a article or anything that involves water treatment in society and carbon intensity can you find the carbon intensity and I will find the society one ?
Yeah sure i was thinking about doing that as well agreed yesterday
I was thinking about doing some work about that at the library tomorrow after class so if you wanna like work on that together?
You
I was thinking about doing some work about that at the library tomorrow after class so if you wanna like work on that together?
Ohh that's perfect yeah I thought it will be better then zoom so we can work on it together and figure it out
I have found some data but im not really sure if its gonna be useful for both imapct on carbon and society


8
Ayaan
You
I was thinking about doing some work about that at the library tomorrow after class so if you wanna like work on that together?
Ohh that's perfect yeah I thought it will be better then zoom so we can work on it together and figure it out
I have found some data but im not really sure if its gonna be useful for both imapct on carbon and society
Ayaan
Ohh that's perfect yeah I thought it will be better then zoom so we can work on it together and figure it out
Sounds like a plan!
You
I have found some data but im not really sure if its gonna be useful for both imapct on carbon and society
We will go over it tomorrow after class and we will see
Alright thank you I will see you tomorrow
See you tomorrow

Ayaan Iqbal (submitted on 29/03/24 20:19):

Personal Contributions:

- My focus was to investigate how water contamination impacts four key areas: health, society, carbon intensity, and the environment.

- I analysed the data provided to us and conducted additional research to understand the significant effects of water contamination on these four lenses.

- This comprehensive approach aligned with the original Project Plan, which aimed to explore the multifaceted impacts of water contamination.

Challenges Faced and Overcome:

- One major challenge I encountered was the lack of sufficient data or resources provided to us, specifically regarding the impact of water contamination on carbon intensity and society.

- To overcome this obstacle, I spoke to my colleague Anas, and we decided to utilize external sources to gather the necessary information and bridge the gaps in our existing data.

- However, finding reliable sources posed another challenge. To ensure the credibility of the information I collected, I relied on reputable platforms such as PubMed and other similar sources.

- I carefully reviewed the references cited in the research papers I found to confirm their reliability and relevance to our project.

- For example, me and Anas discovered a study on PubMed that provided valuable insights into the carbon intensity of wastewater treatment, backed by numerous references from peer-reviewed professionals.

Alignment with Project Plan:

- My personal contributions closely aligned with the original Project Plan's objective of exploring the impact of water contamination on various aspects of life and the environment.

- By focusing on health, society, carbon intensity, and the environment, I adhered to the plan's comprehensive approach.

- However, the plan did not anticipate the challenges of finding sufficient data and resources for certain lenses, particularly carbon intensity and society.

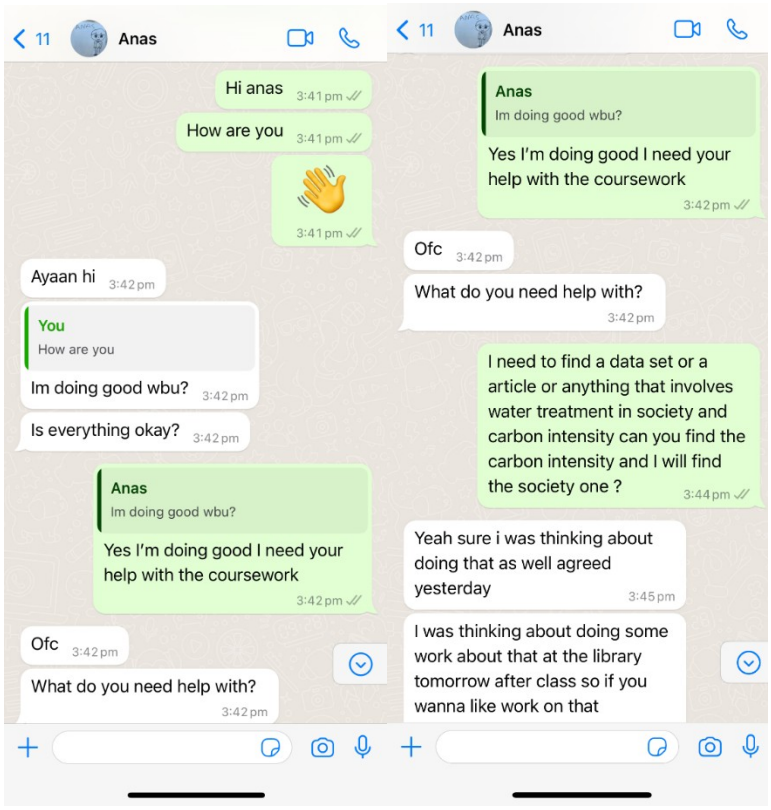
- Despite this deviation, I adapted my approach by utilizing external sources and ensuring their reliability, ultimately fulfilling the project's goals.

Work based learning experiences:

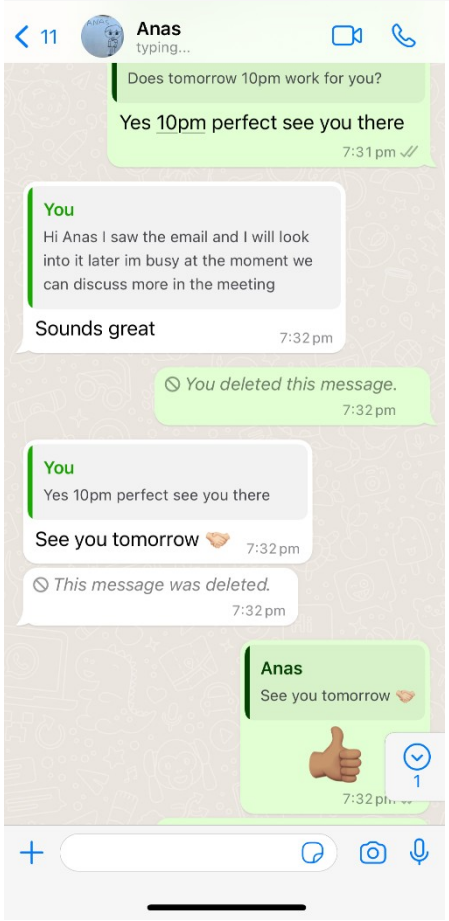
Through my work-based learning experience, I have acquired valuable skills in conducting research, writing analytical pieces, and justifying arguments using data and evidence. I learned to identify relevant information from various sources and present it effectively using visual aids. This experience has enhanced my analytical and writing skills, which are transferable to various academic and professional settings. Additionally, I have developed strong time management abilities, enabling me to deliver high-quality work even under tight deadlines.

Collaborating with my team members:

Collaborating with my team has significantly improved my teamwork and communication skills. I learned to share ideas, listen to others' perspectives, and adapt to different working styles, fostering a positive and productive work environment. Receiving feedback from my peers has been crucial in refining my work and elevating it to a higher standard. It has taught me the importance of being open to constructive criticism and using it as an opportunity for growth. Moreover, peer feedback has highlighted the significance of effective communication and the ability to provide constructive feedback to others.




11
Anas
Yeah sure i was thinking about doing that as well agreed yesterday 3:45 pm
I was thinking about doing some work about that at the library tomorrow after class so if you wanna like work on that together? 3:46 pm
Anas
I was thinking about doing some work about that at the library tomorrow after class so if you wanna like work on that...
Ohh that's perfect yeah I thought it will be better then zoom so we can work on it together and figure it out 3:47 pm
I have found some data but im not really sure if its gonna be useful for both imapct on carbon and society 3:47 pm
You
Ohh that's perfect yeah I thought it will
11
Anas
can work on it together and figure it out 3:47 pm
I have found some data but im not really sure if its gonna be useful for both imapct on carbon and society 3:47 pm
You
Ohh that's perfect yeah I thought it will be better then zoom so we can work on it together and figure it out
Sounds like a plan! 3:47 pm
Anas
I have found some data but im not really sure if its gonna be useful for both imapct on carbon and society
We will go over it tomorrow after class and we will see 3:47 pm
Alright thank you I will see you tomorrow 3:48 pm
See you tomorrow 3:48 pm
11
Anas
Hey ayaan, hope you okay... i have sent you that data i found on Pubmed via email please check it and lemme know 7:30 pm
Also i was thinking of maybe we can do a meeting on zoom and find more data together ! 7:31 pm
Does tomorrow 10pm work for you? 7:31 pm
Hi Anas I saw the email and I will look into it later im busy at the moment we can discuss more in the meeting 7:31 pm
Anas
Does tomorrow 10pm work for you?
Yes 10pm perfect see you there 7:31 pm
You
Hi Anas I saw the email and I will look into it later im busy at the moment we can discuss more in the meeting


11
Anas
typing...
Does tomorrow 10pm work for you?
Yes 10pm perfect see you there 7:31 pm
You
Hi Anas I saw the email and I will look into it later im busy at the moment we can discuss more in the meeting
Sounds great 7:32 pm
You deleted this message. 7:32 pm
You
Yes 10pm perfect see you there
See you tomorrow 7:32 pm
This message was deleted. 7:32 pm
Anas
See you tomorrow
1

References

Olteanu, A. (2018). Country Mapping - ISO, Continent, Region (Version 1.0) [Dataset]. Retrieved from Kaggle: https://www.kaggle.com/datasets/andradaolteanu/country-mapping-iso-continent-

www.destatis.de. (n.d.). *European Statistical System - German Federal Statistical Office*. [online] Available at: https://www.destatis.de/Europa/EN/Methods/ESS/_inhalt.html#173226 [Accessed 27 Feb. 2024].

ec.europa.eu. (n.d.). *Eurostat and the European Statistical System*. [online] Available at: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Eurostat_and_the_European_Statistical_System#European_Statistical_System_.28ESS.29 [Accessed 27 Feb. 2024].

Kohler, J.C. and Bowra, A. (2020). Exploring anti-corruption, transparency, and accountability in the World Health Organization, the United Nations Development Programme, the World Bank Group, and the Global Fund to Fight AIDS, Tuberculosis and Malaria. *Globalization and Health*, 16(1). doi:https://doi.org/10.1186/s12992-020-00629-5.

www.thelancet.com. (n.d.). *About the Global Burden of Disease*. [online] Available at: https://www.thelancet.com/gbd/about.

United Nations (2023). *Water Quality and Wastewater*. [online] UN-Water. Available at: https://www.unwater.org/water-facts/water-quality-and-wastewater.

United Nations (2021). *Water Scarcity*. [online] UN-Water. Available at: https://www.unwater.org/water-facts/water-scarcity.

UN-Water. (n.d.). *UN World Water Development Report 2021*. [online] Available at: https://www.unwater.org/publications/un-world-water-development-report-2021.

Our World in Data. (n.d.). *Annual freshwater withdrawals*. [online] Available at: https://ourworldindata.org/grapher/annual-freshwater-withdrawals?tab=table [Accessed 27 Feb. 2024].
Media Bias/Fact Check. (2017). *Our World In Data - Media Bias/Fact Check*. [online] Available at: https://mediabiasfactcheck.com/our-world-in-data/.

Fida, M., Li, P., Wang, Y. et al. Water Contamination and Human Health Risks in Pakistan: A Review. Expo Health 15, 619–639 (2023). https://doi.org/10.1007/s12403-022-00512-1

Chem. Rev. Metal–Organic Frameworks for the Removal of Emerging Organic Contaminants in Water 120, 16, 8378–8415(2020). https://doi.org/10.1021/acs.chemrev.9b00797

Ishrat Bashir, F. A. Lone, Rouf Ahmad Bhat, Shafat A. Mir, Zubair A. Dar & Shakeel Ahmad Dar, Concerns and Threats of Contamination on Aquatic Ecosystems (2020). https://link.springer.com/chapter/10.1007/978-3-030-35691-0_1

Aldosari, F., Kassem, H.S., Baig, M.B., Muddassir, M. and Mubushar, M., 2017. Impact of sewage on health, economic and social life of rural people in Al-Hair - Kingdom of Saudi Arabia.

Agriculture and Forestry Journal, 1(1), pp.10-17.
https://core.ac.uk/download/pdf/235273806.pdf

Lu, L., Guest, J.S., Peters, C.A., Zhu, X., Rau, G.H., Ren, Z.J., 2018. Wastewater treatment for carbon capture and utilization. Nat. Sustain. 1(12), 750–758.

Li, L., Wang, X., Miao, J., Abulimiti, A., Jing, X., Ren, N., 2022. Carbon neutrality of wastewater treatment - A systematic concept beyond the plant boundary. Environ. Sci. Ecotechnol. (In Press)

IPCC, 2013. Climate Change 2013: the Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, USA.

McCarty, P.L., Bae, J., Kim, J., 2011. Domestic wastewater treatment as a net energy producer--can this be achieved? Environ. Sci. Technol. 45(17), 7100–7106.

Hao, X., Liu, R., Huang, X., 2015. Evaluation of the potential for operating carbon neutral WWTPs in China. Water Res. 87, 424–431.