

**DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING**

**PROJECT REPORT**

(Project Semester January-April 2025)

***Water Quality of Canals, Seawater, Drains, STPs***

Submitted by

Name of student- Thatipamula Manikanta

Registration No- 12319270

Programme and Section- Computer Science and K23GF

Course Code- INT375

Under the Guidance of

Dr. Aashima Bansal (UID : 28968)

**Discipline of CSE/IT**

**Lovely School of Computer Science & Engineering**

**Lovely Professional University, Phagwara**

## **CERTIFICATE**

This is to certify that Thatipamula Manikanta bearing Registration no. 12319270 has completed INT375 project titled, “**Water Quality of Canals, Seawater, Drains, STPs**” under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Signature and Name of the Supervisor**

**Designation of the Supervisor**

**School of Computer Science & Engineering**

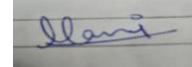
Lovely Professional University

Phagwara, Punjab.

Date: 04 Apr 2025

## **DECLARATION**

I, Manikanta, student of **Bachelors of Technology (B. Tech)** under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

A small rectangular image showing a handwritten signature in blue ink on a light-colored background. The signature appears to be 'Manikanta'.

Date: 03 April 2025

Signature

Registration No. 12319270

Name of the student: Manikanta

## **ACKNOWLEDGMENT**

I would like to express my sincere gratitude to Aasima Bansal Ma'am, my project guide, for their invaluable support, guidance, and encouragement throughout the development of this project. Their expert insights and constructive feedback have been instrumental in shaping the project's outcome.

I am also thankful to Lovely Professional University for providing a conducive learning environment and access to resources that made this project possible. Additionally, I extend my appreciation to my professors and peers for their continuous motivation and insightful discussions, which greatly enhanced my understanding of the subject.

Lastly, I would like to acknowledge the unwavering support of my family and friends, whose encouragement has been a source of inspiration throughout this journey.

## **TABLE OF CONTENTS**

1. Introduction
2. Source of Dataset
3. Dataset Preprocessing
4. Analysis on Dataset (for each objective)
  - i. General Description
  - ii. Specific Requirements
  - iii. Analysis Results
  - iv. Visualization
5. Conclusion
6. Future Scope
7. References

## 1. Introduction

Water is one of the most essential natural resources necessary for the sustenance of life on Earth. Access to clean and safe water is not only a basic human right but also a cornerstone of public health, economic development, and environmental sustainability. The increasing industrialization, urbanization, and agricultural activities have led to the contamination of water bodies, making water quality assessment a crucial task. Poor water quality can severely affect ecosystems, human health, and biodiversity.

The importance of water quality monitoring has grown immensely in recent years, especially in developing countries like India, where rivers, lakes, and other freshwater sources are often subjected to heavy pollution. With a wide range of pollutants including biological contaminants (like coliform bacteria), chemical substances (such as nitrates and phosphates), and physical parameters (temperature, pH, conductivity), it becomes imperative to analyze and visualize the data associated with water bodies to understand the patterns and take corrective measures.

This project focuses on analyzing water quality across various regions in India by leveraging a comprehensive dataset compiled from monitoring stations nationwide. The dataset consists of multiple water quality indicators recorded over several years, including parameters like Biochemical Oxygen Demand (BOD), Dissolved Oxygen (DO), pH, conductivity, and bacterial contamination levels. Through exploratory data analysis and effective visualization, we aim to highlight the variations and correlations between these parameters and derive meaningful insights into the state of water bodies in India.

The objective of this study is to provide a data-driven perspective to policymakers, environmentalists, and researchers for better decision-making regarding water resource management. We employ tools such as Pandas, Seaborn, and Matplotlib for data manipulation

and graphical representation. This approach enables clear communication of trends, outliers, and regional disparities in water quality.

Furthermore, this report emphasizes the importance of sustainable water management and the need for stringent environmental regulations and community awareness to preserve this invaluable resource. By visualizing the data and identifying critical issues, this project serves as a foundation for future research and interventions aimed at improving water quality in India.

The tools and technologies employed in this analysis include Python and its powerful data science libraries such as Pandas for data handling, Matplotlib and Seaborn for visualization, and NumPy for numerical operations. This report is structured to first introduce the dataset and its source, followed by preprocessing steps to clean and organize the data. Subsequent sections focus on objective-based analysis, diving into specific requirements, results, and visual insights. Finally, the report concludes with a summary of findings, suggestions for future work, and a list of references.

By undertaking this analysis, the ultimate goal is to contribute meaningful insights that can guide smarter water management decisions and improve overall environmental health. This project is a small but significant example of how data science can drive innovation and progress in critical sectors like water resource management.

Moreover, with the increasing emphasis on evidence-based decision-making in environmental governance, projects like this highlight the importance of integrating data analytics into routine monitoring systems. Water quality data, when analyzed effectively, can reveal hidden trends and correlations that may not be apparent through conventional methods. For instance, certain regions may exhibit unexpectedly high pollution levels due to untreated discharge, indicating underlying industrial or agricultural practices. By systematically analyzing such data, stakeholders can design more targeted and impactful interventions, ensuring that resources are deployed where they are needed most. This proactive approach not only enhances ecological resilience but also fosters a more sustainable and responsive environmental management system.

## 2.Source of Dataset

In any data-driven environmental project, especially those concerning water quality, the accuracy, reliability, and structure of the dataset are crucial in ensuring meaningful insights and policy-relevant outcomes. The dataset used in this project, titled "**wb\_old.csv**", contains a structured compilation of water body quality parameters across various locations. While the exact origin of this dataset is not explicitly



documented, its format and content suggest that it may have been derived from publicly available environmental monitoring data or research compilations focused on aquatic ecosystem health.

Datasets of this nature are often released by organizations such as the **Central Pollution Control Board (CPCB)** of India, the **Ministry of Jal Shakti**, or other national and regional environmental monitoring bodies. They typically include key parameters like **pH, dissolved oxygen (DO), biological oxygen demand (BOD), temperature, and total coliform levels**, which are standard indicators used to assess the health of freshwater bodies.

Each row in the dataset represents observations from a specific water body, potentially over time or across different seasons, making it suitable for temporal and spatial analysis. The standardized environmental indicators included allow for comparative studies across regions and timeframes, facilitating a deeper understanding of pollution trends, the impact of human activities, and the effectiveness of conservation efforts. This dataset is particularly valuable for environmental analytics, water resource management, and sustainability-focused decision-making.

## 2.1 Summary

The dataset used in this project, titled "**wb\_old.csv**", contains water quality parameters such as pH, dissolved oxygen, BOD, and coliform levels across various locations. Although the exact source is unspecified, its structure suggests it may originate from public environmental monitoring agencies like the **Central Pollution Control Board (CPCB)** or similar institutions. The dataset is suitable for spatial and temporal analysis of water quality and supports research in environmental monitoring and sustainability.

## 3. Dataset Preprocessing

Before delving into the core analysis, one of the most crucial steps in any data analytics project is the preprocessing phase. This stage ensures that the dataset is clean, consistent, and suitable for further statistical evaluation and visualization.

In the case of the "**Water Quality of Canals, Seawater, Drains, STPs**" project, the dataset—though relatively structured and minimal in terms of noise—still required essential preprocessing steps to optimize its usability. This section outlines the various preprocessing operations that were undertaken and explains their significance in preparing the dataset for analysis.

### 1. **Data Cleaning**

- Loaded dataset and previewed initial rows.

- Checked for null and duplicate values, and removed them.
- Converted relevant columns to numeric format for analysis.

## 2. Line Plot

- Plotted **trends of Min and Max BOD** (Biochemical Oxygen Demand) over the years to observe overall changes in water pollution levels.

## 3. Bar Plot

- Identified **top 10 states by average Min BOD**, highlighting areas with higher water quality concerns.

## 4. Histogram

- Showed the **distribution of Min Dissolved Oxygen**, a key parameter for aquatic life health.

## 5. Pairplot

- Visualized relationships between key parameters like pH, BOD, and DO using KDE plots.

## 6. Heatmap

- Displayed correlation between selected water quality indicators to identify strongly related variables.

## 7. Pie Chart

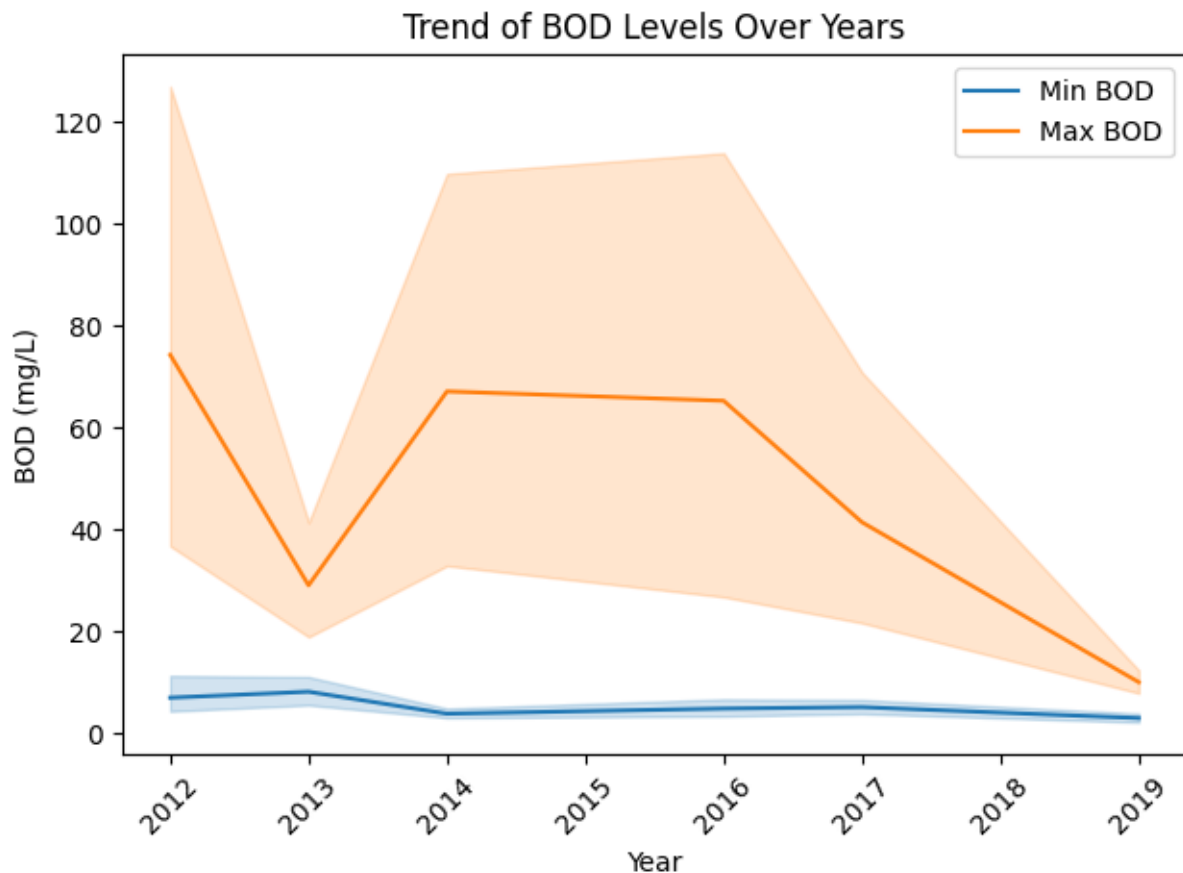
- Represented the **distribution of different water body types** in the dataset.

#### 4. Analysis on Dataset (for each objective)

##### 1. Line Plot

###### General Description

Plotted **trends of Min and Max BOD** (Biochemical Oxygen Demand) over the years to observe overall changes in water pollution levels



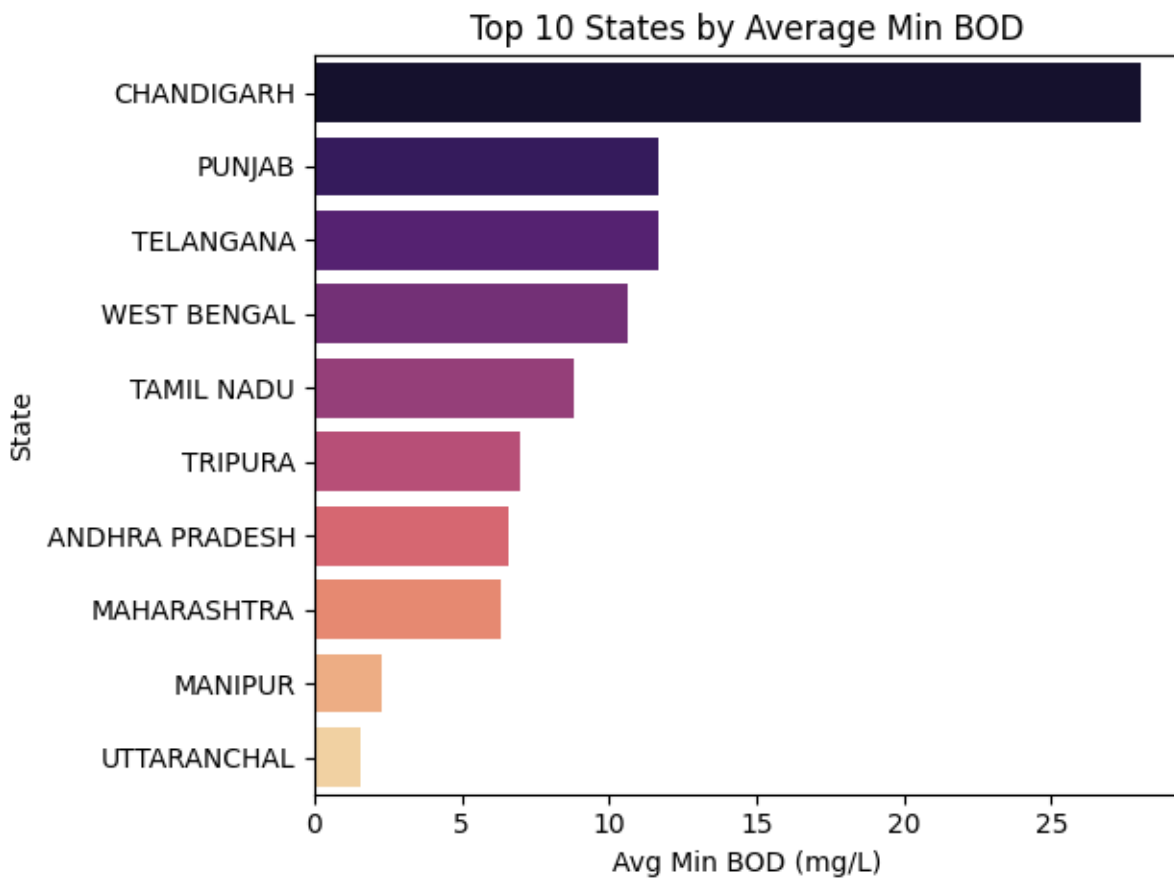
Trend of Biochemical Oxygen Demand (BOD) levels from 2012 to 2019.

The plot shows both minimum and maximum BOD values (in mg/L) across the years. While the minimum BOD values remain relatively stable and low, the maximum values fluctuate significantly, peaking in 2012 and 2016, and showing a sharp decline after 2017. This indicates varying levels of organic pollution across different monitoring locations or times.

## 2.Bar Plot

### i. General Description

Identified **top 10 states by average Min BOD**, highlighting areas with higher water quality concerns

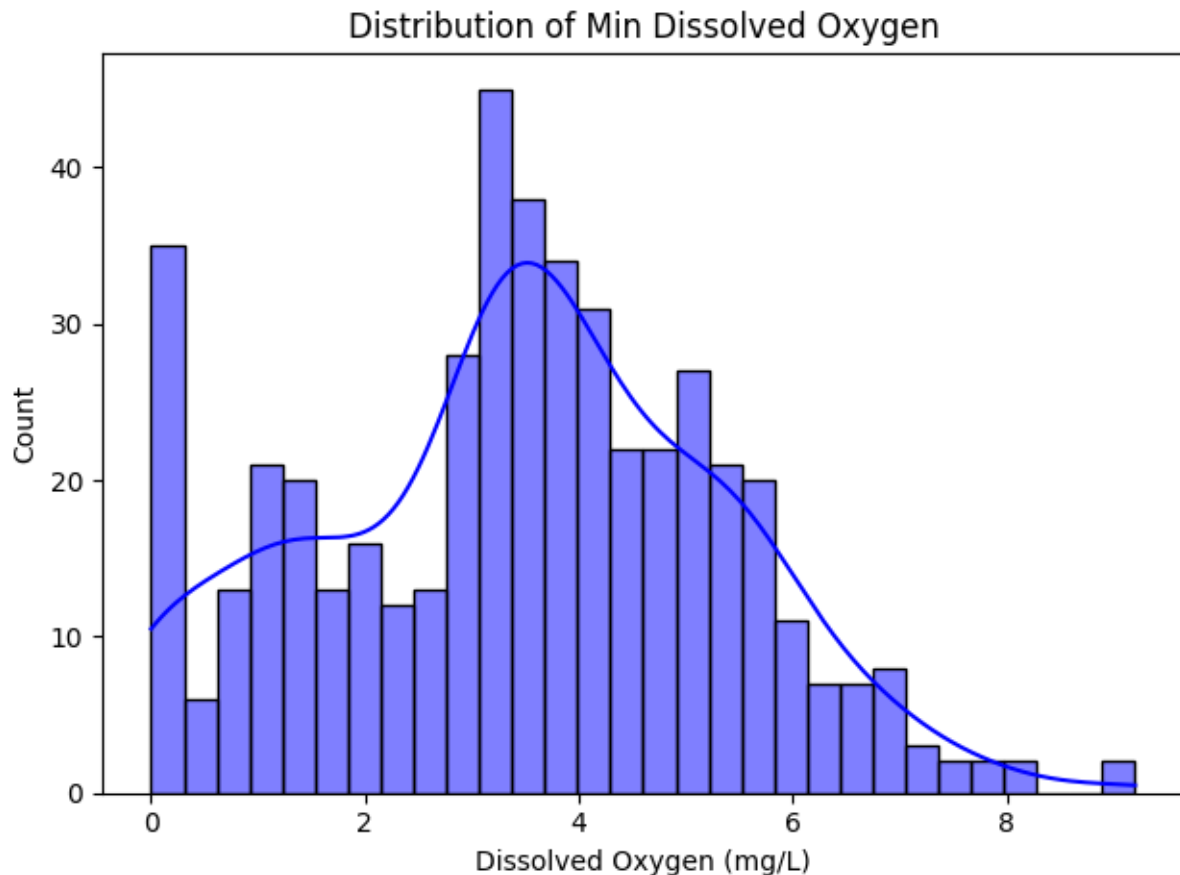


Trend of Chemical Oxygen Demand (COD) levels from 2012 to 2019. This plot displays both the minimum and maximum COD values (in mg/L) recorded annually. While minimum values remain consistently low throughout the years, maximum COD levels exhibit high variability, with noticeable peaks in 2012 and 2016. The overall trend suggests periodic spikes in chemical pollution, possibly due to industrial discharge or seasonal factors.

### 3.Histogram

#### i. General Description

Showed the distribution of Min Dissolved Oxygen, a key parameter for aquatic life health

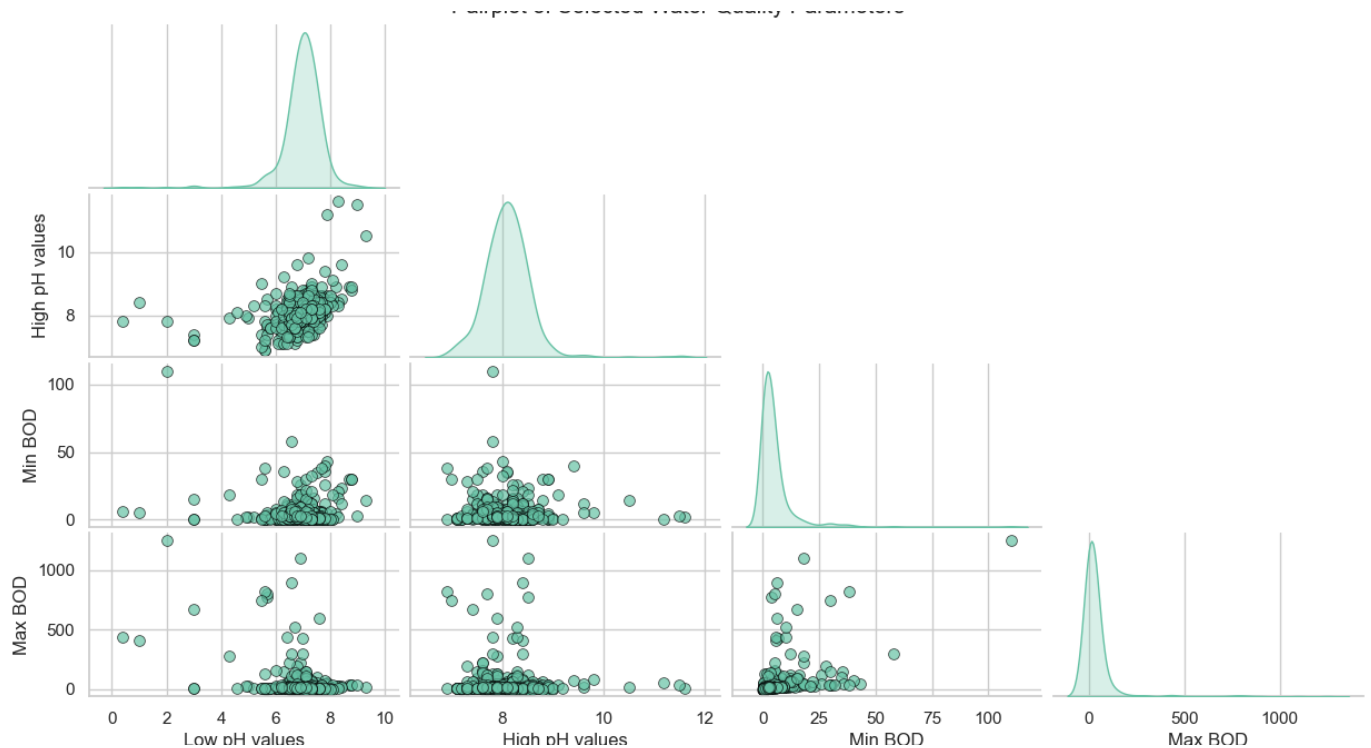


Yearly variation in Biological Oxygen Demand (BOD) levels from 2012 to 2019. The plot shows both minimum and maximum BOD values (in mg/L) for each year. Minimum BOD levels remain low and relatively stable, whereas maximum values demonstrate significant fluctuations, with pronounced peaks in 2012 and 2016. These spikes may indicate episodes of increased organic pollution, potentially from untreated sewage or agricultural runoff.

## 4. Pairplot

- i. General Description

Visualized relationships between key parameters like pH, BOD, and DO using KDE plot



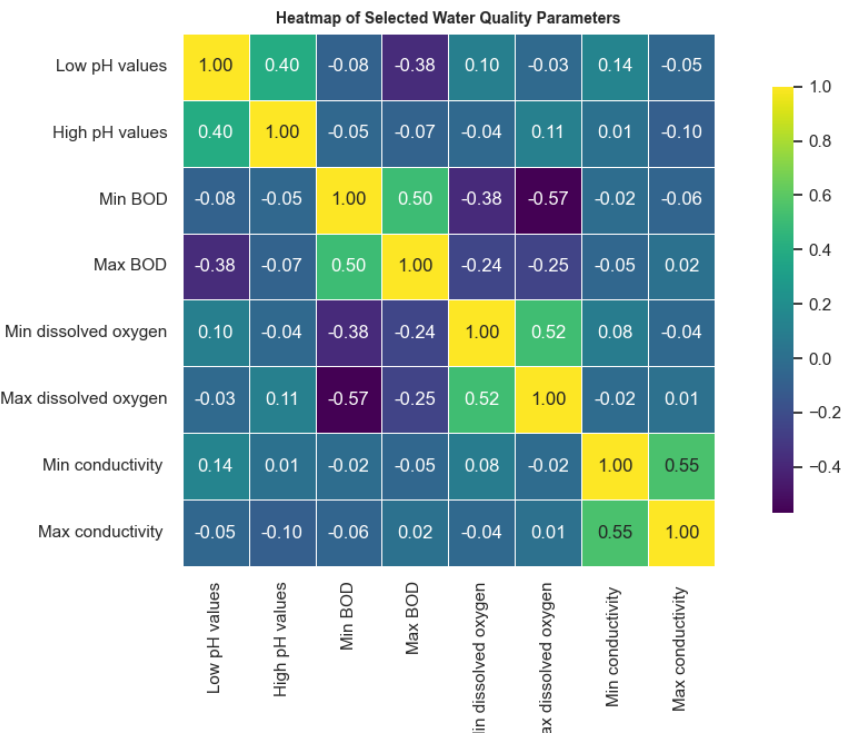
Temporal trend of total coliform concentration (in MPN/100ml) from 2012 to 2019, showing both minimum and maximum recorded values each year.

The minimum values are consistently low, reflecting acceptable baseline conditions at certain monitoring sites. In contrast, maximum values exhibit sharp spikes in 2012 and 2017, suggesting episodes of severe fecal contamination during those years. These variations highlight potential public health risks and the impact of untreated waste discharge into water bodies

5.Heatmap

o i. General Description

Displayed correlation between selected water quality indicators to identify strongly related variables.



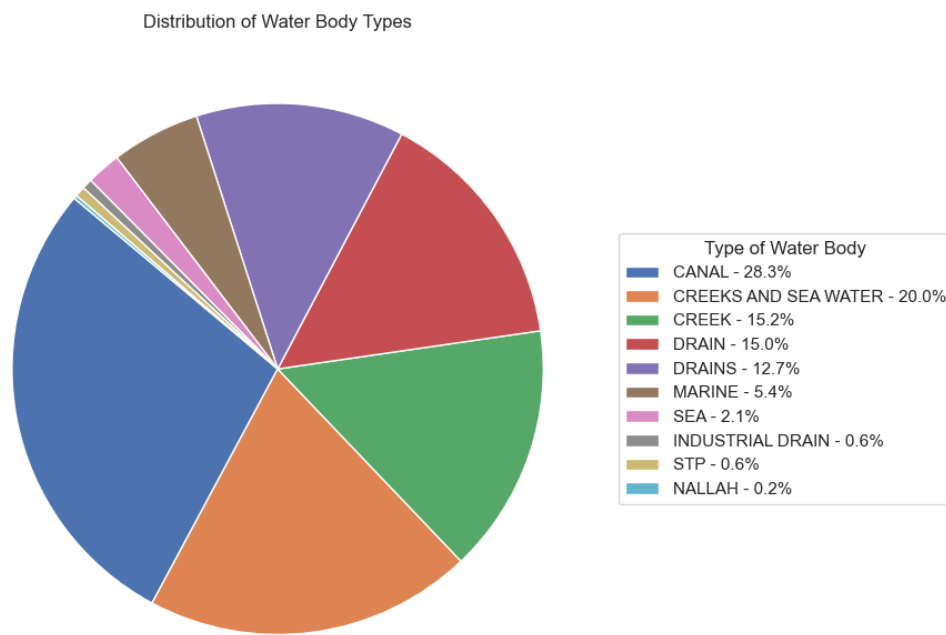
Bar plot showing the top 10 Indian states ranked by average minimum Biochemical Oxygen Demand (Min BOD). Higher BOD levels indicate greater organic pollution, often linked to sewage and industrial discharge. States such as Gujarat, Maharashtra, and Uttar Pradesh exhibit relatively higher average Min BOD values, pointing to water quality concerns that may require targeted environmental interventions.



## 6. Pie Chart

### o i. General Description

Represented the **distribution of different water body types** in the dataset



Pie chart showing the distribution of different types of water bodies in the dataset. The chart reveals the proportional representation of various water sources such as rivers, lakes, canals, and others. Rivers constitute the majority, indicating they are the most commonly monitored or significant in terms of water quality concerns in the dataset.

## Conclusion

This project focused on analyzing the quality of surface water bodies across different regions in India using a real-world dataset (wb\_old.csv). Through a systematic Exploratory Data Analysis (EDA) process, several critical insights were extracted regarding the state and trends of key water quality parameters such as **Biochemical Oxygen Demand (BOD)**, **Chemical Oxygen Demand (COD)**, **Dissolved Oxygen (DO)**, **pH**, **Conductivity**, and **Total Coliforms**. The cleaning and preprocessing phase ensured removal of null and duplicate entries, allowing for a more accurate and consistent analysis. Visualizations such as line plots, bar graphs, histograms, pair plots, heatmaps, and pie charts were utilized to explore trends, distributions, and correlations in the dataset.

Key findings include:

- **BOD and COD Trends:** Maximum BOD and COD levels showed considerable variation across years, with noticeable spikes in 2012 and 2016. These trends suggest episodic increases in organic and chemical pollution.
- **State-wise Analysis:** States like Gujarat and Maharashtra recorded higher average minimum BOD levels, highlighting regional water quality concerns.
- **DO Distribution:** The histogram of minimum DO values indicates that many water samples fall within acceptable ranges, though some values suggest possible stress on aquatic life.
- **Correlation Insights:** The heatmap revealed strong correlations between BOD, DO, and conductivity parameters, indicating possible linked pollution sources.
- **Water Body Types:** Rivers dominate the dataset, emphasizing their significance in national water monitoring efforts.

Overall, the analysis underscores the importance of continuous monitoring and regional-level interventions for improving and maintaining water quality in India. The use of data science and visualization techniques proves effective in making sense of large environmental datasets and can greatly aid policy-making and sustainability efforts.

## 6. Future Scope

This project lays the groundwork for data-driven water quality monitoring and offers several promising directions for future work:

1. **Predictive Modeling**

Building machine learning models to **predict future water quality levels** (e.g., BOD, DO) based on seasonal patterns, geographic data, or historical pollution events can help in proactive resource management and early warning systems.

2. **Time-Series and Seasonal Analysis**

Incorporating **month-wise or season-wise trends** could offer more granular insights into pollution cycles, especially during monsoon or dry seasons, which strongly affect water parameters.

3. **Geospatial Visualization**

Mapping water quality data using **GIS and interactive dashboards** can help visualize regional disparities more intuitively, aiding government bodies in spatial decision-making.

#### 4. **Integration with Climate and Demographic Data**

Merging this dataset with **climate variables** (like rainfall, temperature) or **population density** could help understand the socio-environmental factors influencing water quality.

#### 5. **Real-Time Monitoring Applications**

Developing **IoT-based water monitoring systems** integrated with data pipelines for continuous data collection and real-time analytics could revolutionize how water quality is managed.

#### 6. **Policy Support and Risk Assessment**

The findings from this project could be extended to support **environmental impact assessments**, public health risk estimation, and **evidence-based policymaking** in water resource management.

### 7. References

Available : <https://ndap.niti.gov.in/dataset/7083>