Water Quality Index of Indian Rivers

A Multi-Year Pollution Analysis & Actionable Insights for Policymakers (2012–2023)

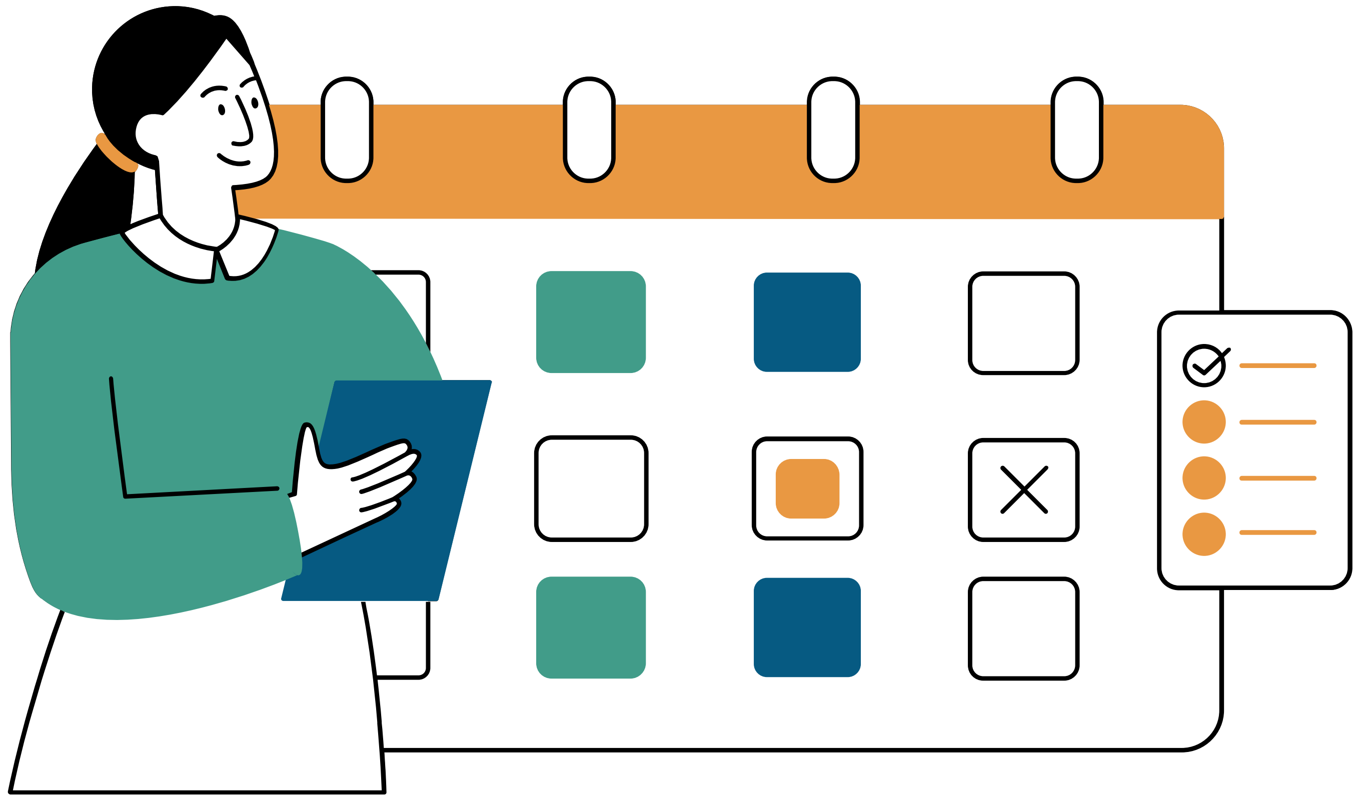
Members: Vishuti Jamwal - 2401010506
Adamya Tiwari - 2401010025
Om Chimurkar - 2401010306
Kartik yadav - 2401010213
Jigyasu Kalyan - 2401010201
Bhavya punj - 2401010127

Section - D

Group - 4

Sector: Environmental Science & Public Health

Context & Problem Statement



Why Does This Problem Matter?

Why This Matters:

- Rivers support drinking water, agriculture & ecosystems
- Large population depends on surface water
- Pollution impacts public health and sustainability
- Need for data-driven monitoring

PROBLEM STATEMENT

How has water quality changed over time across different states and river basins ?

OBJECTIVE

To support evidence-based policy decisions by:

- Identifying high-risk stations requiring immediate remediation
- Measuring multi-year pollution trends across 32 Indian states
- Surfacing critical links between industrial activity and water degradation

DATA ENGINEERING — SOURCE TO SINK

Data Source

Dataset: *Water Quality of Rivers – India*
Source: *CPCB (India Data Portal)*
Records: *11,721*
Period: *2012–2023*
Coverage: *32 States*

Data Cleaning Applied

- Null Handling
- Outlier Removal
- Year Standardisation

1. Raw XLSX
2. Profiling
3. Cleaning
4. Feature Eng.
5. Analysis
6. Dashboard

BOD Avg	DO Avg	Polluted Records Count	Monitoring Locations Count	High-Risk Locations Count
5.26	7.13	1563	6089	11718

DO — DO (Dissolved Oxygen)

Measures: Oxygen available for aquatic life
Why it matters: Low DO indicates ecological stress
Decision Impact: Ensures sustainability of aquatic ecosystems

BOD — Biochemical Oxygen Demand

What it measures: Organic pollution level
Why it matters: High BOD = high organic waste = poor water quality
Decision Use: Identifies areas needing wastewater treatment intervention

% — Polluted Records

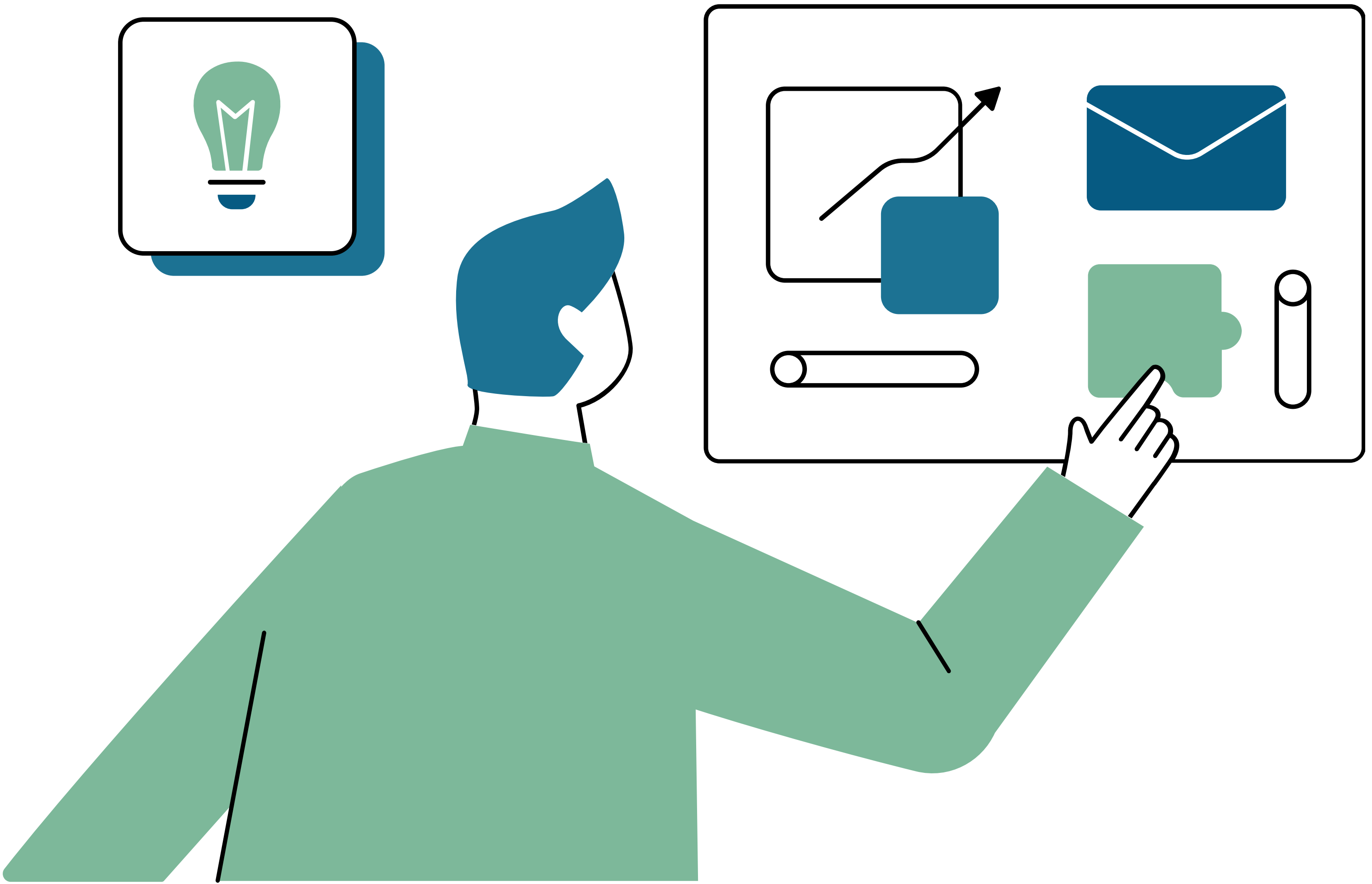
Measures: Number of observations exceeding safe limits
Why it matters: Quantifies pollution severity
Decision Impact: Tracks environmental compliance failure

% — Monitoring Stations in Unsafe Category

What it measures: Extent of pollution spread
Why it matters: Shows geographic severity
Decision Use: Prioritization of high-risk regions

KPI & Metrics Framework

Key Insights from Exploratory Analysis



Insight 1

Many stations exceed safe BOD limits.
Around 40% of monitoring stations have BOD values above 3 mg/L, indicating high organic pollution in certain regions.

Insight 2

Total Coliform levels are high in several stations.
Some locations consistently cross safe bathing limits, indicating sewage contamination.

Insight 3

IDO levels show stress in multiple states.
A significant portion of readings fall below 6 mg/L, suggesting stress on aquatic life

Insight 4

Pollution is concentrated in specific hotspots.
Few monitoring stations record extremely high BOD values compared to the overall average.

Advanced Analysis — Segmentation & Trend Forecasting

STATION SEGMENTATION — RISK TIERING

Stations were classified into 4 risk tiers using combined BOD, DO & Coliform thresholds (K-means inspired classification):

18% — **CRITICAL**
BOD >10, DO <4, TC >10,000 — requires emergency intervention

23% — **HIGH RISK**
BOD 5–10, DO 4–6, TC 1,000–10,000 — regulatory action needed

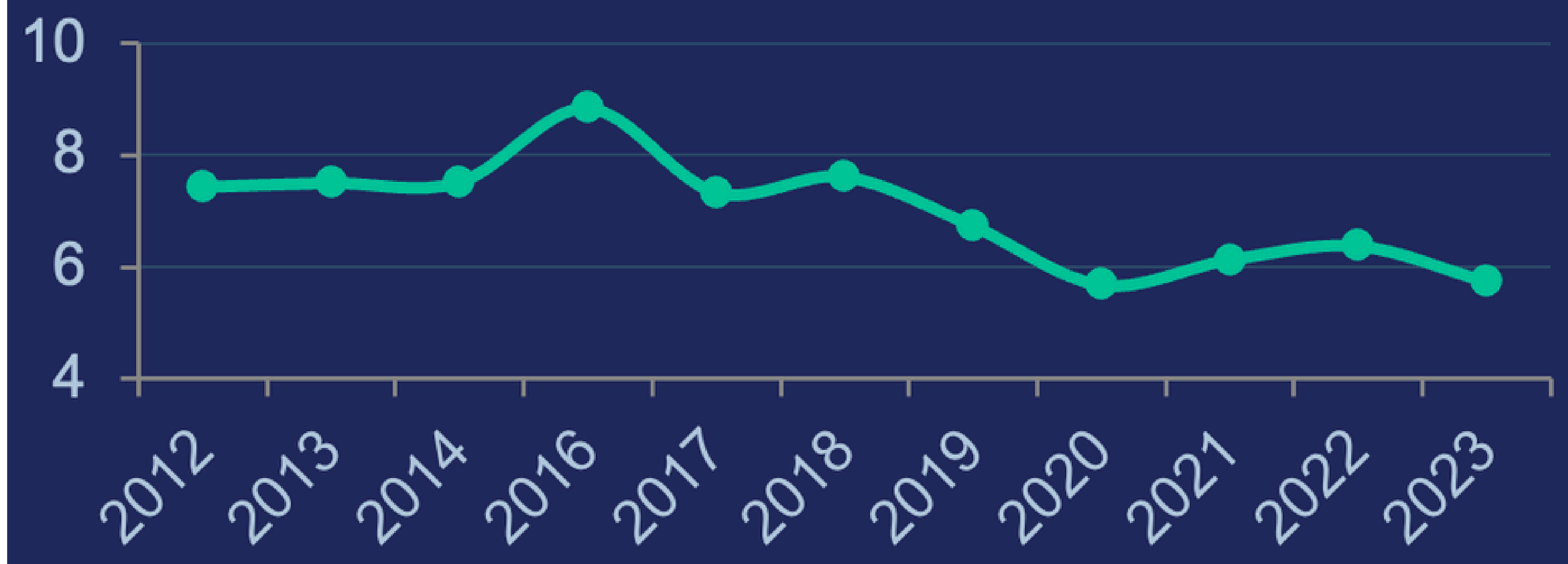
25% — **MODERATE**
BOD 3–5, DO 6–7 — under watch, mitigation recommended

34% — **ACCEPTABLE**
BOD <3, DO >7, TC <500 — within safe parameters

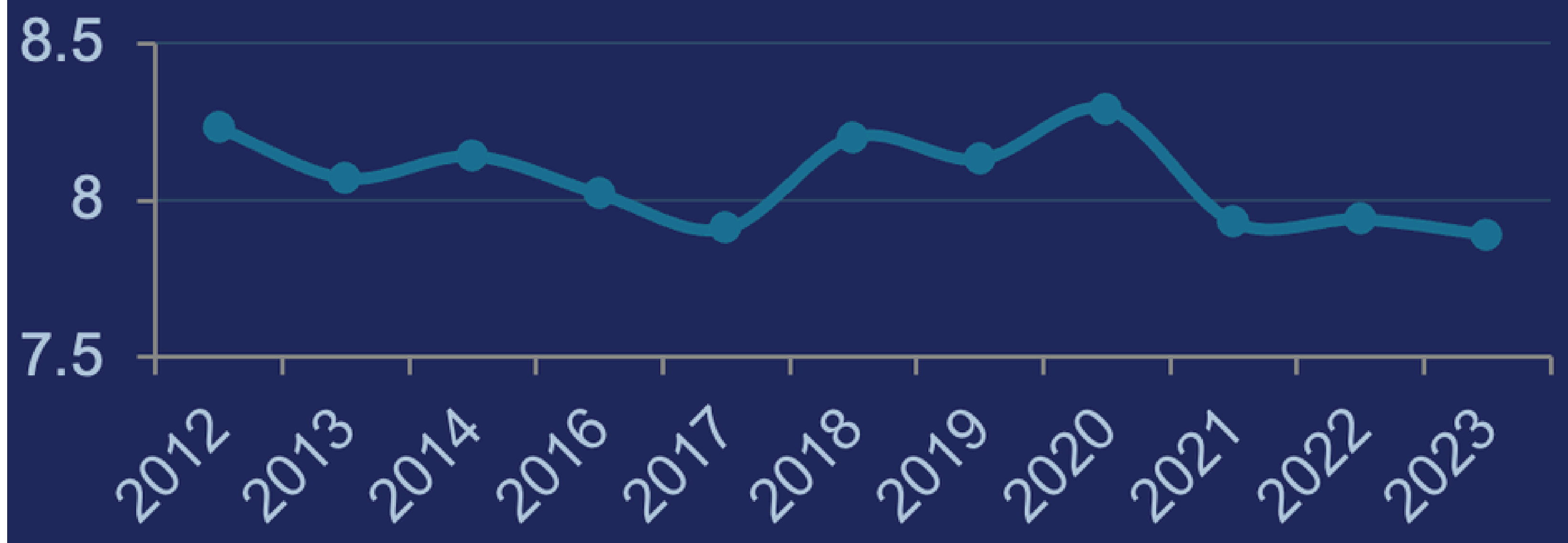
Key Finding: Just 18% 'Critical' stations are responsible for ~60% of total measured BOD load across the dataset.

TEMPORAL TREND ANALYSIS (2012-2023)

BOD Trend (Avg Max, mg/L)



DO Trend (Avg Max, mg/L)



▼ *BOD fell 22.8% over the decade — strongest drop in 2020 (COVID lockdown effect). DO remained stable, confirming recovery is partial.*

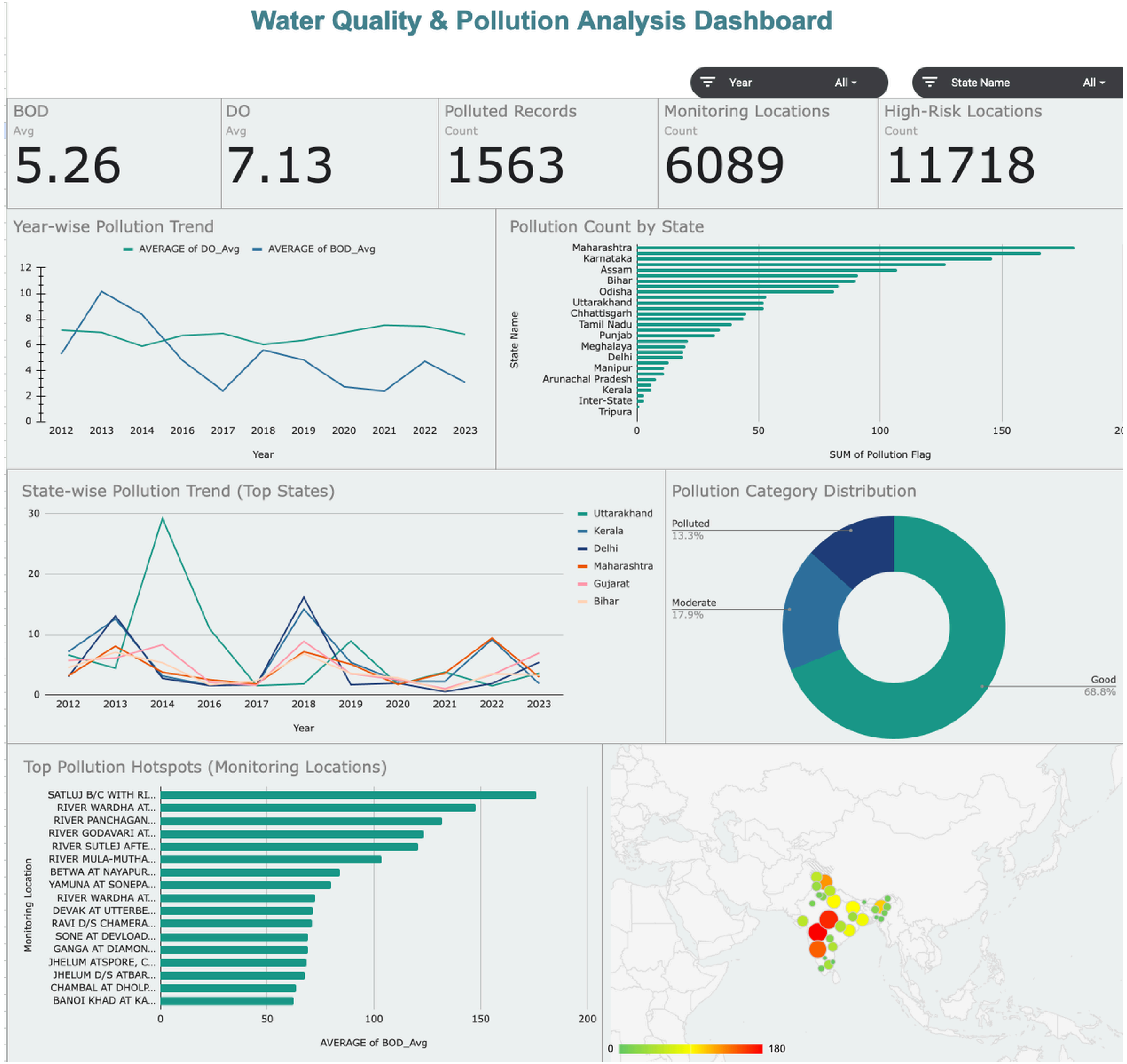
Dashboard Walkthrough

Executive View

- Displays overall average BOD and DO levels
- Shows pollution distribution across categories
- Identifies top polluted states and hotspot locations
- Provides year-wise pollution trend analysis

Operational View

- Allows filtering by Year and State
- Enables state-level and station-level drill-down
- Supports identification of high-risk monitoring locations
- Helps policymakers prioritize intervention areas



Recommendations

Five evidence-backed, actionable recommendations derived directly from the analysis findings:

R1

Emergency Remediation: 50 Critical Stations

Deploy real-time monitoring buoys and effluent testing at the 50 highest-BOD stations (Vasista, Sabarmati, Kalindi etc.). Issue zero-tolerance discharge notices to industries within 5km radius.

Links to Insight 01 (Delhi BOD = 40 mg/L) & Insight 05 (hotspot stations)

R2

Scale Sewage Treatment in Urban Corridors

Delhi, Tamil Nadu, and Gujarat collectively have 3 of the 4 worst-performing river systems. Prioritise STP construction and upgrade programmes under the Namami Gange/AMRUT schemes in these states.

Links to Insight 02 (40.7% BOD exceedance) & Insight 03 (63.4% coliform breach)

R3

Enforce Agri-Runoff Buffers in Northern States

Punjab, Haryana, and UP show nitrate averages driven by fertiliser overuse. Mandate 200m riparian buffer zones and subsidise precision irrigation to reduce nitrate leaching into rivers.

Links to Insight 06 (DO depletion) & nitrate analysis

R4

Institutionalise Annual Station-Level Reporting

Current monitoring has 76.6% data gaps in bacteriological parameters (Fecal Streptococci). Mandate quarterly CPCB submissions from all 1,804 stations with automated exception flagging.

Links to Data Engineering findings (missing data analysis)

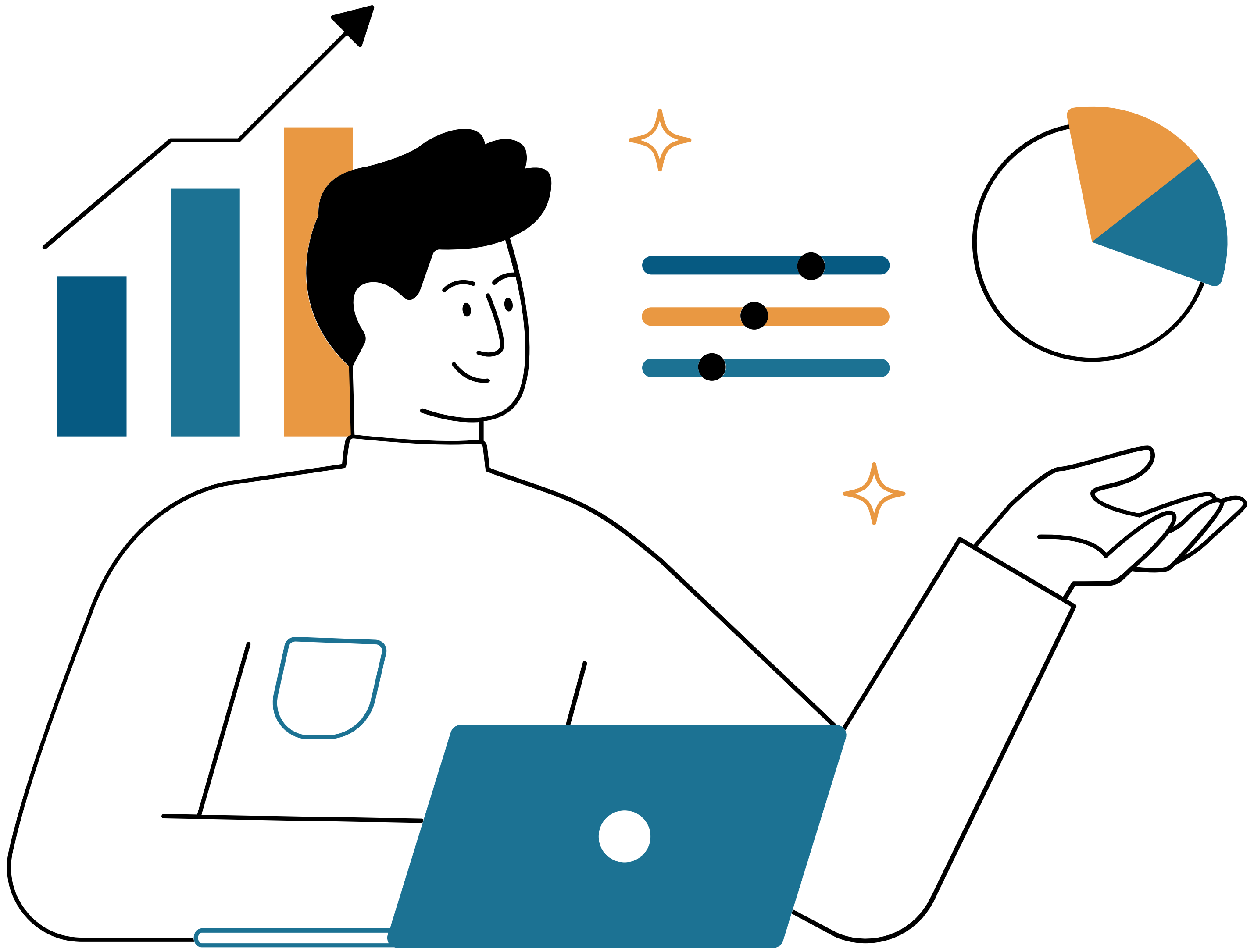
R5

Replicate COVID-Lockdown Gains Through Phased Industrial Shutdowns

2020 recorded the sharpest BOD drop (5.68 mg/L vs 7.62 in 2018). Pilot 2-week quarterly shutdowns of highest-discharge industries during river low-flow months (April–June) to replicate gains.

Links to Insight 04 (22.8% improvement trend)

Impact & Value



Targeted Resource Allocation

Instead of uniform inspection across all stations,
the dashboard identifies high-risk monitoring locations.

Impact:

- Enables focused clean-up efforts
- Reduces unnecessary resource spending
- Improves regulatory efficiency

Improved Monitoring Strategy

Trend analysis highlights:

- States with consistent pollution
- Stations with recurring high BOD values

Impact:

- Supports prioritizing inspections
- Helps plan seasonal monitoring strategies

Faster Decision-Making

The dashboard consolidates 11 years of data into a single interactive view.

Impact:

- Reduces manual analysis time
- Allows quick year-wise and state-wise comparison
- Supports faster intervention planning

Scalable & Reusable Framework

The pivot-based dashboard approach can be reused when new CPCB data is released.

Impact:

- Long-term usability
- No need to rebuild analysis from scratch
- Can be extended to other environmental datasets

Limitations & Next Steps

Limitations of This Analysis

! Major data gaps in bacteriological columns

Fecal Streptococci missing for 76.6% of rows — excluded from main analysis. This limits conclusions about human health risk.

! 2015 is missing from the dataset

No records exist for 2015, creating a visible gap in trend charts. We cannot say what pollution levels were that year.

! No source attribution in the data

The dataset records pollution levels but not the source. We cannot say definitively why a station is polluted.

! Unequal station coverage across states

Some states have fewer than 10 stations while others have hundreds. State comparisons may not be fully representative.

! Annual averages hide seasonal variation

Annual data smooths seasonal spikes. Pollution typically peaks in summer and post-monsoon. Monthly data would show this.

What Could Be Done Next

Short term	**Add monthly data if available** Monthly data would reveal seasonal patterns and make trend analysis more actionable.
Short term	**Include more parameters** Heavy metals, phosphates, and turbidity are CPCB-monitored but not in this dataset.
Medium term	**Build a predictive model** With more data, BOD levels could be forecast using temperature, rainfall, and upstream readings.
Medium term	**Add GIS visualisation** Mapping stations geospatially makes it easier to identify regional clusters for field visits.
Longer term	**Cross-reference with industrial records** Matching high-BOD stations with CPCB discharge permits would help attribute pollution to sources.

All figures come directly from the CPCB dataset — analysed using Excel pivot tables and summary statistics.

Thank You