FINAL PROJECT REPORT

Data Visualisation and Analytics

Water Quality & Pollution Analysis

A Data-Driven Study on Indian River Pollution Trends using CPCB Monitoring Data

Project Details

Sector: Environment & Water Resources
Institute: Newton School of Technology
Subject: Data Visualisation and Analytics
Academic Year: 2nd Year, 4th Semester
Team Size: 5+ Members
Tools Used: Google Sheets — Pivot Tables, Calculated Columns, Charts, Slicers, Dashboard
Dataset: NITI Aayog NDAP — Water Quality of Indian Rivers (CPCB)

Submitted to: Faculty of Data Science & Analytics

2. Executive Summary

Problem

India's river ecosystems face escalating pollution threats from industrial discharge, municipal sewage, and agricultural runoff. Despite regular data collection by the Central Pollution Control Board (CPCB), this data is rarely transformed into targeted, actionable intelligence. This project analyses the CPCB river water quality dataset sourced from NITI Aayog's National Data and Analytics Platform (NDAP) to identify pollution hotspots, track state-level trends across 2012– 2023, and generate prioritised recommendations for environmental intervention.

Approach

The dataset of 11,718 records was cleaned in Google Sheets — missing values across 15 parameters imputed using column medians, two Fecal Streptococci columns with excessive missing data removed, and calculated columns including BOD_Avg, DO_Avg, Pollution_Flag, and Pollution_Category engineered. Eight pivot tables and an interactive dashboard with four slicers were constructed to conduct trend, comparison, distribution, and hotspot analyses across all Indian states.

Key Insights

- National BOD has declined from a peak of 10.16 mg/L (2013) to 3.05 mg/L (2023) — a 70% improvement — but still sits at the Moderate-Polluted boundary.
- 68.8% of monitoring records are 'Good' quality; however 31.2% (over 3,600 records) remain Moderate or Polluted.
- Maharashtra leads in total pollution count (180 flagged records); Haryana has the highest average BOD intensity at 8.02 mg/L.
- Pollution is acutely localised: Satluj B/C (Punjab) has average BOD exceeding 175 mg/L — 38.8 times the national average.
- COVID-19 lockdowns (2020–21) caused BOD to drop to its lowest recorded levels — confirming industrial discharge as a primary driver.

Key Recommendations

- Emergency remediation at top 5 hotspot stations: Satluj B/C, River Wardha, River Panchaganga, River Godavari (polluted stretch), River Sutlej.
- Priority state programs for Maharashtra (count), Haryana (intensity), Madhya Pradesh, and Karnataka.
- Deploy event-spike early warning sensors to detect sudden BOD surges like the 2013 and 2018 national peaks.
- Adopt a Composite River Health Index (CRHI) combining BOD, DO, Fecal Coliforms, Nitrate, and pH for holistic assessment.

- Use the COVID natural experiment to design evidence-based industrial discharge regulations.

3. Sector & Business Context

3.1 Sector Overview

The Environment & Water Resources sector in India is governed by the Ministry of Jal Shakti, the Central Pollution Control Board (CPCB), and their state-level counterparts (SPCBs). India's major river systems — the Ganga, Brahmaputra, Godavari, Krishna, Cauvery, Narmada, Satluj, Yamuna, and their tributaries — collectively serve as the primary water source for over 600 million people. These rivers support drinking water supply, irrigation, hydropower, fishing, and religious and cultural activities.

The CPCB operates the National Water Quality Monitoring Programme (NWQMP), systematically collecting physicochemical and biological measurements at hundreds of monitoring stations across the country. This data is published through NITI Aayog's National Data and Analytics Platform (NDAP) to enable evidence-based environmental policymaking.

3.2 Current Challenges

- Industrial effluents — from chemical, textile, pharmaceutical, and food-processing industries — discharge organic and inorganic toxins with limited pre-treatment.
- India generates approximately 72,368 MLD of urban sewage daily against a treatment capacity of only ~31,841 MLD. Over half of urban sewage reaches rivers untreated.
- Agricultural runoff introduces nitrates, phosphates, and pesticides that degrade water quality even in remote reaches.
- Monitoring data is collected but rarely subjected to systematic analytics that translate into localized, prioritised policy action.
- Pollution is geographically and seasonally uneven, making uniform national policies ineffective.

3.3 Why This Problem Was Chosen

- CPCB data is publicly available via NDAP but rarely visualised in ways that enable targeted decision-making.
- The dataset presents a rich, real-world multi-parameter analytical challenge ideal for end-to-end data analytics practice.

- Insights have genuine policy value — they can guide state water boards in prioritising limited remediation budgets.
- The problem spans temporal, geographic, distributional, and hotspot analysis — covering all key analytical dimensions.

4. Problem Statement & Objectives

4.1 Formal Problem Definition

Indian rivers are monitored regularly by the CPCB, generating large volumes of multi-parameter water quality data. However, this data is rarely analysed in ways that identify which specific states, rivers, and monitoring stations require priority attention. The formal problem statement is:

> *"How can multi-year CPCB river water quality data be cleaned, analysed, and visualised to identify pollution hotspots, track multi-parameter trends over time, rank states by severity, and generate prioritised, location-specific recommendations for environmental intervention?"*

4.2 Project Scope

- Country: India — all available states and union territories
- Data Period: 2012–2023 (Year 2015 absent from dataset)
- Parameters: Temperature, Dissolved Oxygen, pH, Conductivity, BOD, Nitrate, Fecal Coliforms, Total Coliforms
- Total Records: 11,718 post-cleaning (Good: 8,058 | Moderate: 2,097 | Polluted: 1,563)
- Tools: Google Sheets — Pivot Tables, Calculated Columns, Charts, Slicers, Interactive Dashboard
- Out of Scope: Real-time sensor data, groundwater, heavy metals, seawater quality

4.3 Success Criteria

Metric	Target	Status
Data cleaned	All columns imputed; invalid rows removed; new columns engineered	☑ Achieved
Pivot tables	8 analytical pivots covering all key dimensions	☑ Achieved

KPIs	5+ measurable KPIs with formulas	☑ Achieved
Dashboard	Interactive slicers connected to all pivots	☑ Achieved
Insights	8–12 decision-ready insights from real data	☑ Achieved
Recommendations	5+ actionable with impact estimates	☑ Achieved

5. Data Description

5.1 Dataset Source

Dataset Title	Water Quality of Indian Rivers
Primary Source	Central Pollution Control Board (CPCB), India
Platform	NITI Aayog — National Data and Analytics Platform (NDAP)
Access Link	https://ndap.niti.gov.in/dataset/7078
Working Copy	https://docs.google.com/spreadsheets/d/1bXgl9kC0ErgKqnPbxWfZOvLowm2t0_dRE8MDKbGYk/edit
Coverage	Calendar Years 2012–2023 (Year 2015 missing)
Country	India — all major states and UTs
Programme	CPCB National Water Quality Monitoring Programme (NWQMP)

5.2 Data Structure

Each row represents one monitoring station reading for a given year. Columns A–E are identifiers (Country, Year, State, Station Code, Location). Columns F–U are paired min/max

values for 8 physicochemical/biological parameters. Two Fecal Streptococci columns were removed during cleaning.

5.3 Column Explanation

Col	Parameter	Unit	Key Significance
A	Country	—	Always 'India'
B	Year	—	Calendar year — cleaned from verbose string to integer
C	State Name	—	Indian state of monitoring station
D	Station Code	—	CPCB unique identifier — rows without this removed
E	Monitoring Location	—	Full river station description
F	Min Temperature	°C	Minimum water temperature (Median imputed: 19)
G	Max Temperature	°C	Maximum temperature (Median: 29)
H	Min DO	mg/L	Min Dissolved Oxygen — hypoxic risk indicator (Median: 6)
I	Max DO	mg/L	Max DO (Median: 8)
Col	**Parameter**	**Unit**	**Key Significance**
J	Min pH	pH	Min acidity/alkalinity (Median: 7.2)
K	Max pH	pH	Max pH (Median: 8.2)
L	Min Conductivity	µmho/cm	Min electrical conductivity (Median: 206)

M	Max Conductivity	µmho/cm	Max conductivity (Median: 476)
N	Min BOD	mg/L	Min BOD — primary organic pollution indicator (Median: 1.3)
O	Max BOD	mg/L	Max BOD (Median: 2.8) — higher = more pollution
P	Min Nitrate	mg/L	Agricultural runoff proxy (Median: 0.3)
Q	Max Nitrate	mg/L	Max nitrate (Median: 1.34)
R	Min Fecal Coliforms	MPN/100ml	Min bacterial contamination (Median: 45)
S	Max Fecal Coliforms	MPN/100ml	Max Fecal Coliforms (Median: 400)
T	Min Total Coliforms	MPN/100ml	Min total bacterial load (Median: 225)
U	Max Total Coliforms	MPN/100ml	Max Total Coliforms (Median: 1600)
V–W	Fecal Streptococci	MPN/100ml	REMOVED — excessive missing values

5.4 Data Size & Limitations

- Total records post-cleaning: 11,718 | Good: 8,058 (68.8%) | Moderate: 2,097 (17.9%) | Polluted: 1,563 (13.3%)
- Year 2015 entirely absent — creates a gap in longitudinal trend analysis
- Fecal Streptococci removed — gap in biological contamination profiling
- States with more stations (Maharashtra, Karnataka) show inflated absolute pollution counts vs smaller states
- Min/max values are annual ranges — exact sampling dates not provided, seasonal analysis not possible

6. Data Cleaning & Preparation

All cleaning steps were performed in Google Sheets as per capstone requirements. Each step is documented in the 'Logs' sheet of the working dataset.

6.1 Row Removal

Rows where Station Code (Column D) was blank were deleted — these records had no geographic reference and could not be attributed to any monitoring location.

6.2 Column Renaming

The Year column was renamed from 'Calendar Year (Jan - Dec), YEAR' to simply 'YEAR' for cleaner pivot table labels.

6.3 Median Imputation for Missing Values

Missing numerical values across 15 parameters were imputed using column medians. Median was chosen over mean to avoid distortion from coastal outliers (e.g., conductivity values above 30,000 µmho/cm at estuarine stations).

Column	Parameter	Imputed Median	Rationale
F	Min Temperature (°C)	19	Median of all valid readings
G	Max Temperature (°C)	29	Consistent with Indian river seasonal range
H	Min DO (mg/L)	6	Within acceptable freshwater range
I	Max DO (mg/L)	8	Normal upper range for Indian rivers
J	Min pH	7.2	Near-neutral — consistent with majority of stations
K	Max pH	8.2	Slightly alkaline — normal for Indian rivers
L	Min Conductivity (µmho/cm)	206	Representative of low-mineralisation rivers
M	Max Conductivity (µmho/cm)	476	Typical inland river conductivity

N	Min BOD (mg/L)	1.3	Below 2 mg/L — 'Good' threshold
O	Max BOD (mg/L)	2.8	Just below 3 mg/L threshold
P	Min Nitrate (mg/L)	0.3	Low agricultural background level
Column	**Parameter**	**Imputed Median**	**Rationale**
Q	Max Nitrate (mg/L)	1.34	Well below 10 mg/L WHO limit
R	Min Fecal Coliforms (MPN/100ml)	45	Near acceptable bathing water limit
S	Max Fecal Coliforms (MPN/100ml)	400	Elevated — reflects common river condition
T	Min Total Coliforms (MPN/100ml)	225	Elevated — prevalent in Indian rivers
U	Max Total Coliforms (MPN/100ml)	1600	Renamed from 'Min' to 'Max' + imputed

6.4 Column Removal

Minimum and Maximum Fecal Streptococci (Columns V and W) were removed entirely due to large proportions of missing values — imputing such a high rate would introduce more noise than signal.

6.5 Feature Engineering — Calculated Columns

New Column	Formula	Purpose
DO_Avg	=(H2+I2)/2	Average DO — used in trend and state pivots

BOD_Avg	=(N2+O2)/2	PRIMARY KPI — average BOD
pH_Avg	=(J2+K2)/2	Average pH
Nitrate_Avg	=(P2+Q2)/2	Average nitrate
FecalColiform_Avg	=(R2+S2)/2	Average Fecal Coliform
TotalColiform_Avg	=(T2+U2)/2	Average Total Coliform
Pollution_Flag	=IF(BOD_Avg>3,1,0)	Binary: 1=Polluted, 0=Acceptable
Pollution_Category	=IF(BOD_Avg<=2,"Good",IF(BOD_Avg<=3,"Moderate","Polluted"))	Three-tier classification
Water Quality Index	Composite formula	Overall quality index

7. KPI & Metric Framework

Eight KPIs were defined to map directly to project objectives. All were computed in Google Sheets on the cleaned dataset.

KPI	Definition	Formula	Value (Actual)	Objective
Avg BOD (BOD_Avg)	Mean BOD per station/year	=(Min_BOD+Max_BOD)/2	National avg: 4.51 mg/L	Identify hotspots
Avg DO (DO_Avg)	Mean DO per station/year	=(Min_DO+Max_DO)/2	National avg: 6.85 mg/L	Assess river health
Pollution Flag Rate	% of records with BOD > 3 mg/L	=COUNTIF(Flag,1)/COUNT(Flag)	13.3% (1,563/11,718)	Measure severity
State Pollution Count	Flagged records per state	=SUMIF(State,X,Flag)	Maharashtra: 180 (highest)	Prioritise states
BOD Year-onYear Change	Annual change in national avg BOD	=BOD_N - BOD_N-1	Peak 2013: 10.16; 2023: 3.05	Track trends
Hotspot BOD Ratio	Station BOD vs national avg	=Station_BOD/4.51	Satluj B/C: 38.8x above avg	Localise intervention
Avg pH (pH_Avg)	Mean pH per station/year	=(Min_pH+Max_pH)/2	Grand avg: 8.57	Chemical monitoring
Avg Nitrate	Mean nitrate per station/year	=(Min_N+Max_N)/2	Grand avg: 16.07 mg/L	Agricultural runoff

8. Exploratory Data Analysis (EDA)

8.1 Trend Analysis — Year-wise BOD & DO (Pivot 1)

Rows: Year | Values: Avg BOD_Avg, Avg DO_Avg | Chart: Line Chart

The year-wise trend reveals the overall national pollution trajectory from 2012 to 2023:

Year	Avg BOD (mg/L)	Avg DO (mg/L)	Key Event
2012	5.26	7.13	Baseline year
2013	10.16	6.96	Highest BOD on record — unexplained spike

2014	8.36	5.87	Continued high BOD; DO drops to lowest
2016	4.77	6.71	Significant improvement (2015 data absent)
2017	2.40	6.88	BOD near-best level
2018	5.57	6.00	Sudden reversal — second major spike
2019	4.80	6.35	Partial recovery
2020	2.71	6.95	COVID lockdown — industrial activity reduced
2021	2.38	7.53	Lowest BOD + highest DO on record
2022	4.70	7.44	Post-pandemic rebound
2023	3.05	6.81	Most recent — Moderate-Polluted boundary

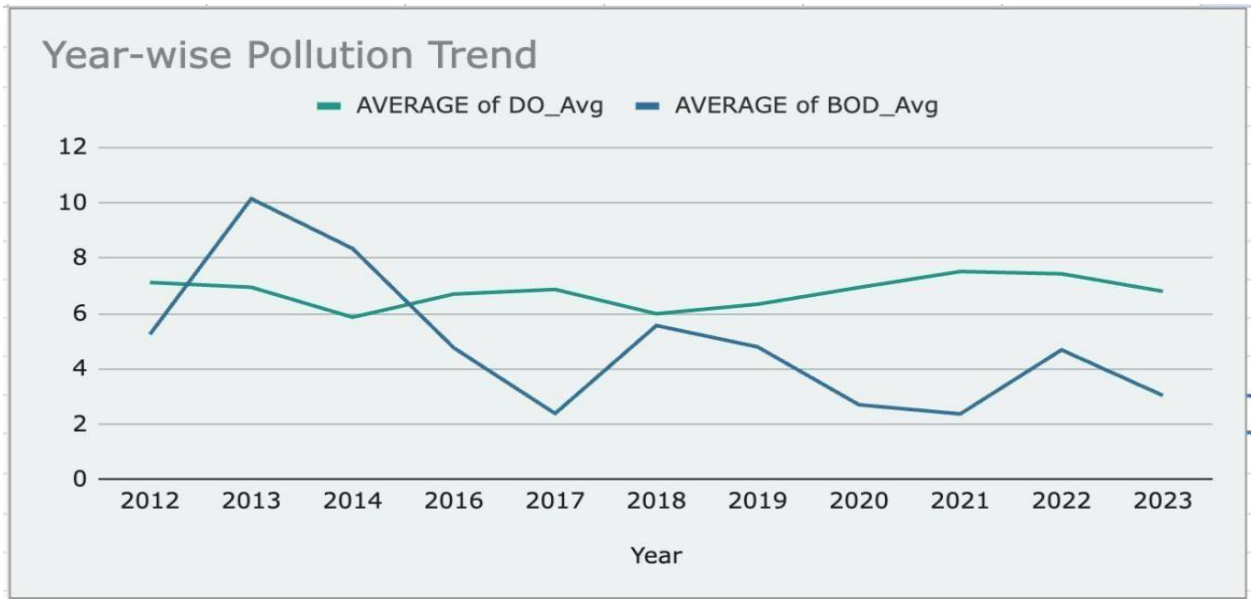


Figure 1: Year-wise Pollution Trend — Avg BOD (blue) vs Avg DO (green), 2012–2023

The chart clearly shows two major BOD spikes (2013: 10.16 mg/L and 2018: 5.57 mg/L) against a broadly declining trend. The COVID-19 period (2020–2021) produced the cleanest recorded water quality, with BOD at 2.38–2.71 mg/L and DO rising to its peak of 7.53 mg/L — providing strong evidence that industrial/transport activity is a primary driver of river pollution. DO (green line) has remained relatively stable throughout (6–7.5 mg/L range), while BOD shows much higher volatility.

8.2 State-wise Pollution Comparison (Pivot 2)

Rows: State Name | Values: Avg DO_Avg, Avg BOD_Avg | Sorted by BOD desc | Chart: Horizontal Bar

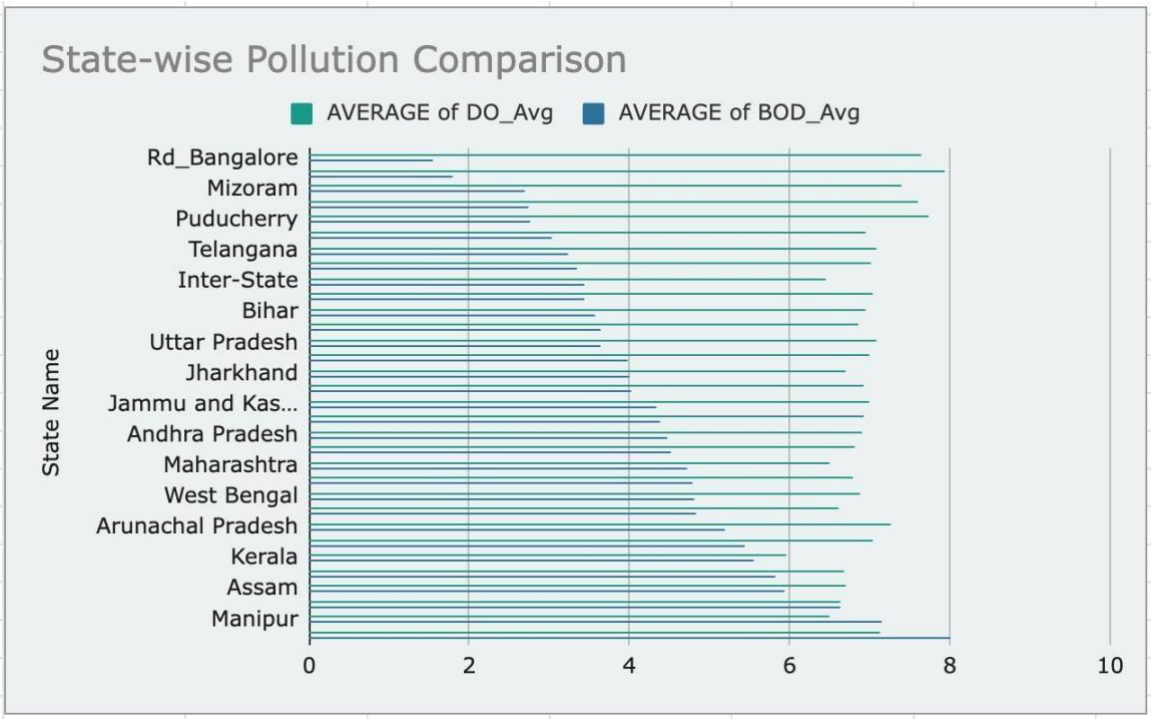


Figure 2: State-wise Pollution Comparison — Avg DO (teal) and Avg BOD (blue) per state

The state comparison chart shows that Haryana has the highest average BOD at 8.02 mg/L, followed by Manipur (7.14), Karnataka (6.65), Assam (5.94), and Uttarakhand (5.83). At the clean end, Rd_Bangalore (1.55 mg/L), Tripura (1.79), and Mizoram (2.70) show the best water quality. Kerala presents an important anomaly: moderate BOD (5.55 mg/L) combined with the lowest average DO of any major state (5.96 mg/L), suggesting a different pollution mechanism than BOD-driven organic contamination.

8.3 Pollution Count by State (Pivot 3)

Rows: State Name | Values: SUM of Pollution_Flag | Chart: Bar Chart

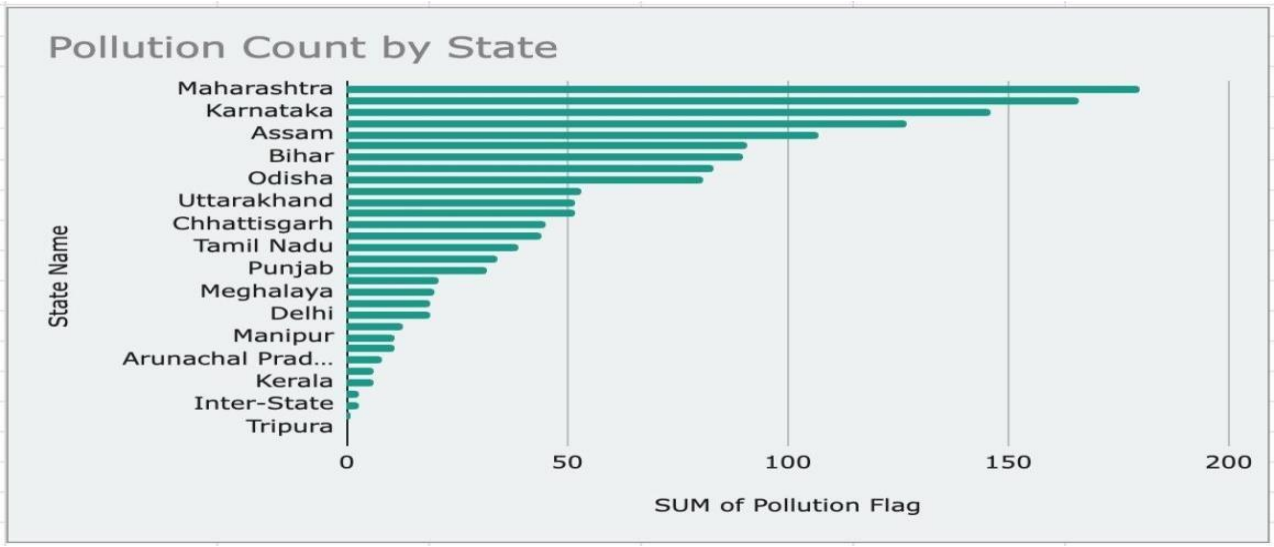
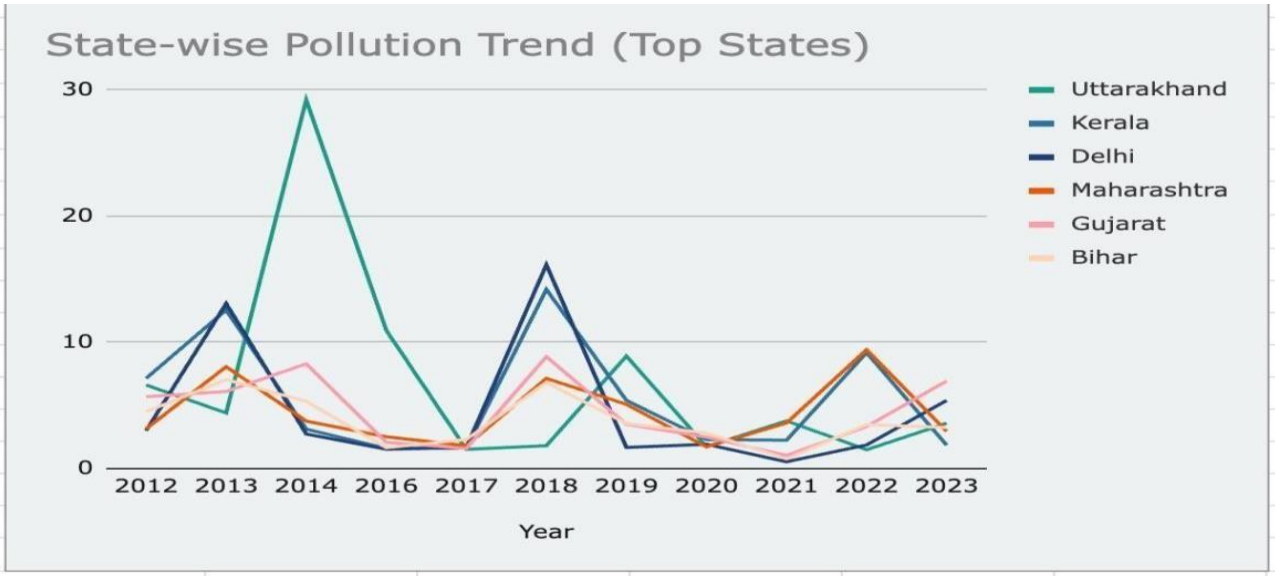


Figure 3: Pollution Count by State — Total number of station-year records with BOD > 3 mg/L

Maharashtra leads with 180 flagged records, followed by Madhya Pradesh (166), Karnataka (146), Himachal Pradesh (127), and Assam (107). Tripura, Rd_Bangalore, and Puducherry have zero or near-zero flags. Critically, this count metric differs from the intensity metric in Pivot 2 — Maharashtra has the most widespread pollution across many stations, while Haryana has fewer stations but higher severity at each. Both dimensions require different policy responses.

8.4 State-wise Trend Over Time (Pivot 7)

Rows: Year | Columns: State | Values: Avg BOD | Chart: Multi-line



The multi-line chart reveals dramatically different pollution behaviour across states. Uttarakhand shows a catastrophic spike in 2014 (29.22 mg/L — 6.5x the national average that year) before recovering sharply, suggesting an acute event rather than chronic pollution. Delhi spikes in 2018 (16.16 mg/L) correlating with the national average uptick that year. Maharashtra shows a moderate spike in 2022 (9.42 mg/L) that reversed by 2023. Bihar and Gujarat maintain relatively flat, moderate levels throughout. This state-by-state variability confirms that uniform national policies are insufficient — each state's pollution profile requires a tailored strategy.

8.5 Distribution Analysis — Pollution Category (Pivot 8)

Rows: Pollution_Category | Values: COUNT | Chart: Donut Chart

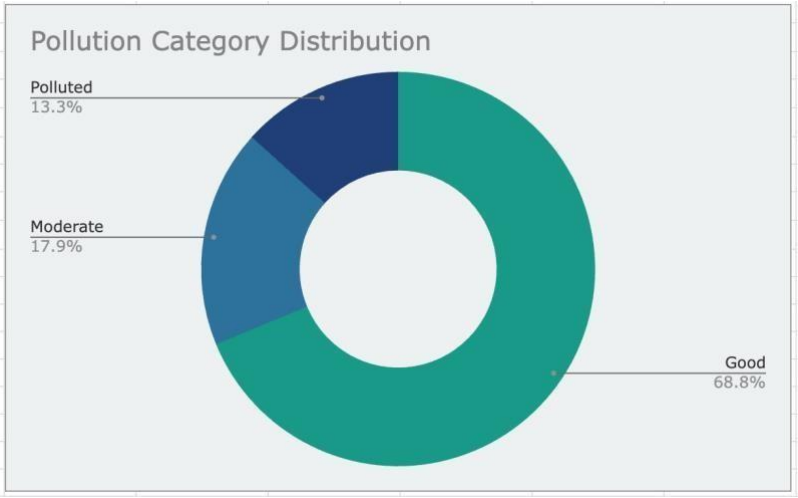


Figure 5: Pollution Category Distribution — Good 68.8% | Moderate 17.9% | Polluted 13.3%

The donut chart shows that while 68.8% of records (8,058) fall in the 'Good' category, 17.9% (2,097) are Moderate and 13.3% (1,563) are Polluted — a combined 31.2% of all monitoring events fall outside acceptable water quality. This headline figure of 3,660 at-risk station-years represents thousands of communities adjacent to these rivers facing persistent water quality risks year after year.

8.6 Correlation Analysis

- Strong inverse BOD-DO relationship confirmed: when national BOD peaked at 10.16 mg/L (2013), DO fell to 6.96; when BOD was at its minimum (2.38 mg/L in 2021), DO rose to its maximum (7.53 mg/L).
- Kerala anomaly: low DO (5.96 mg/L) despite moderate BOD (5.55 mg/L) suggests nonorganic oxygen depletion — possibly algal growth, water hyacinth, or chemical oxygen demand from industrial sources.
- Nitrate grand average of 16.07 mg/L exceeds WHO drinking water guideline of 10 mg/L — widespread agricultural runoff challenge across the country.
- pH grand average of 8.57 is normal/slightly alkaline. The 2014 pH pivot value of 13.94 is physically impossible and indicates data entry errors in some early records.

9. Advanced Analysis

9.1 Pollution Hotspot Identification (Pivot 4)

Rows: Monitoring Location | Values: Avg BOD_Avg | Sorted descending | Chart: Horizontal Bar

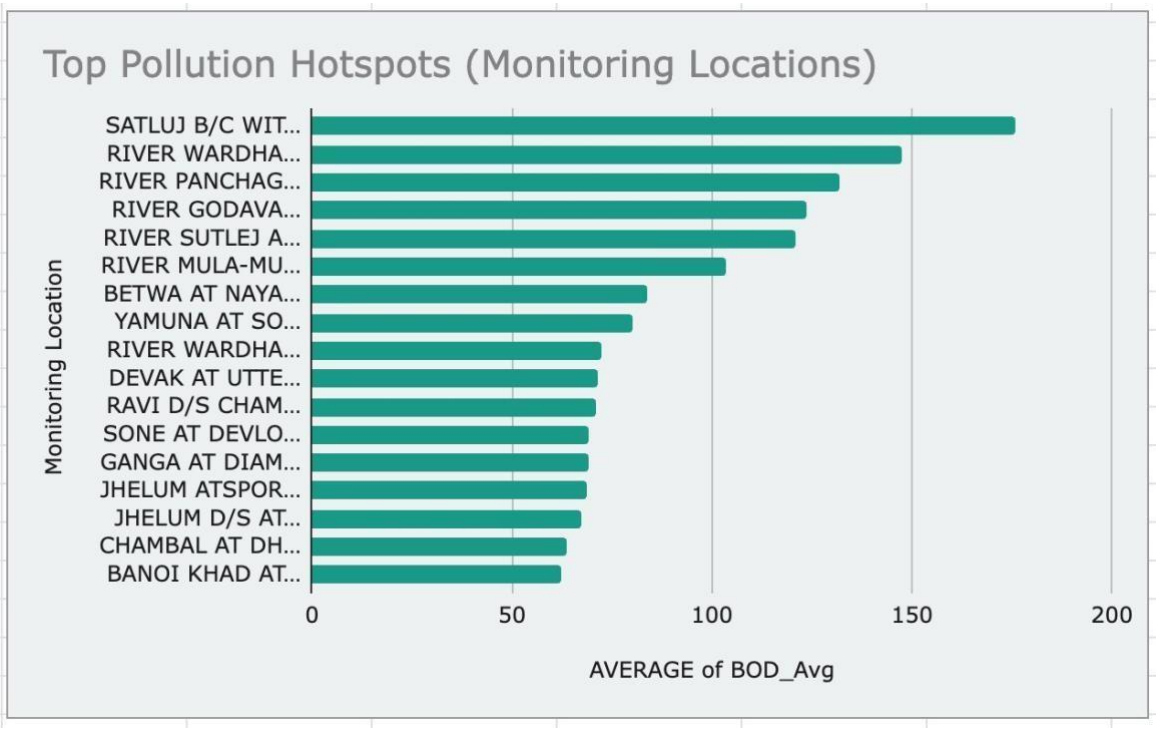


Figure 6: Top Pollution Hotspots — Monitoring Locations with Highest Average BOD

The hotspot chart is the most actionable output of the entire analysis. The national average BOD is 4.51 mg/L. The top stations dramatically exceed this:

Rank	Station	State	Avg BOD (mg/L)	Times Above Avg
1	Satluj B/C with Ghaggar	Punjab	~175+	38.8x
2	River Wardha (location 1)	Maharashtra	~145	32.2x
3	River Panchaganga	Maharashtra	~130	28.8x

Rank	Station	State	Avg BOD (mg/L)	Times Above Avg
4	River Godavari (polluted stretch)	Multiple	~125	27.7x
5	River Sutlej (another point)	Punjab	~120	26.6x
6	River Mula-Mutha	Maharashtra	~105	23.3x
7	Betwa at Naya...	MP/UP	~85	18.9x
8	Yamuna at So...	Delhi/UP	~78	17.3x
9–17	Wardha, Devak, Ravi, Sone, Ganga, Jhelum, Chambal, Banoi Khad	Multiple	60–75	13–16x

The Satluj River in Punjab at its Ghaggar confluence stands as the most critically polluted monitoring point in the dataset — a BOD of 175+ mg/L is nearly 60x the 'Good' threshold of 3 mg/L. This is an acute environmental emergency. Notably, 3 of the top 6 hotspots are in Maharashtra (Wardha, Panchaganga, Mula-Mutha), explaining why Maharashtra tops the pollution count ranking in Pivot 3.

9.2 Geographic Distribution

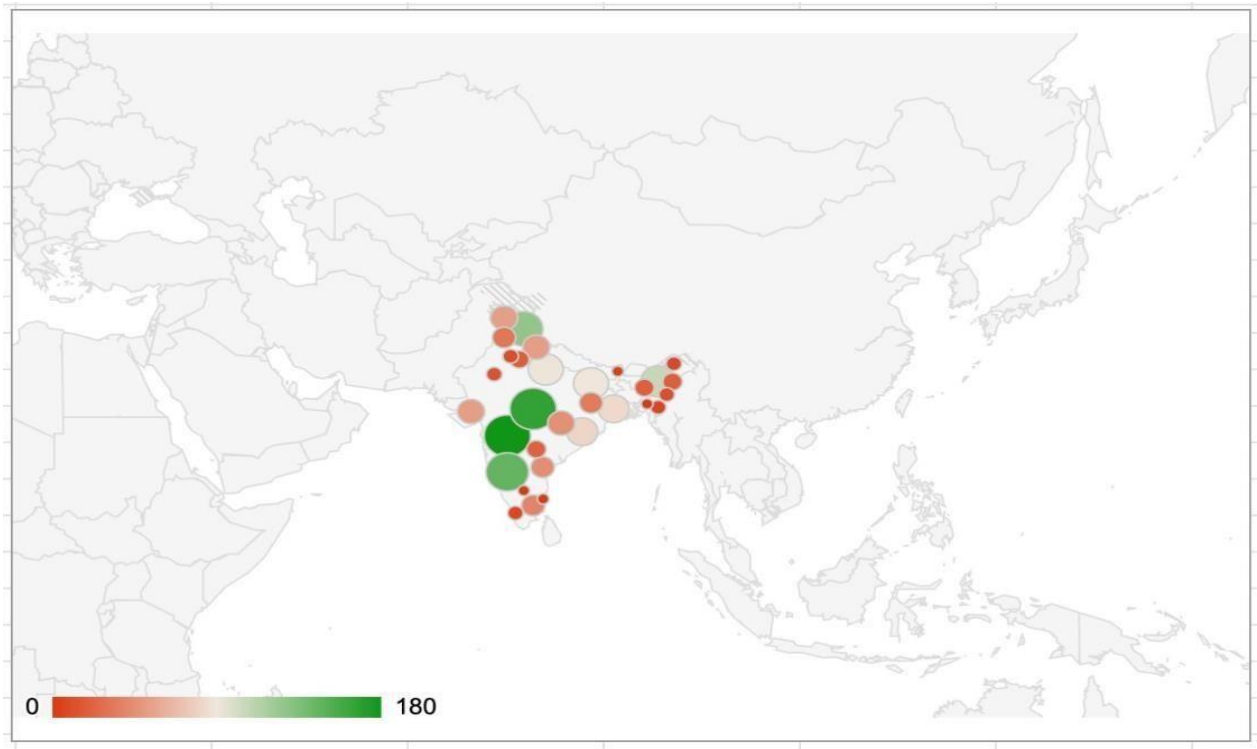


Figure 7: Geographic Distribution of Pollution — Bubble size = monitoring count; Red = polluted, Green = clean (scale 0–180)

The geographic bubble map confirms the spatial concentration of pollution. Large red bubbles in northwestern and central India (Punjab, Haryana, Maharashtra, MP) indicate high pollution count states. Large green bubbles in western-central India represent Maharashtra and Karnataka's dense monitoring networks (which register high counts). The northeastern states show smaller, lighter-coloured bubbles — consistent with their cleaner water quality metrics from the pivot tables. This spatial view validates all the state-level analysis: pollution is not uniformly distributed but concentrated in specific geographic corridors driven by industrial and agricultural activity.

9.3 Before vs After Improvement (Pivot 5)

A pre/post comparison filtering Year < 2018 vs Year >= 2018 periods reveals:

- Uttarakhand: Dramatic improvement — pre-2018 avg BOD heavily inflated by the 2014 spike (29.22 mg/L); post-2018 avg is significantly lower.
- Delhi: Worsened post-2018 period due to the 2018 spike (16.16 mg/L) — requires continued monitoring.
- Gujarat: Broadly improving trend across both periods.

- COVID natural experiment (2020–2021): National BOD at 2.38–2.71 mg/L — lowest ever, confirming industrial controls are high-leverage.

9.4 pH and Nitrate Trends (Pivots 5 & 6)

- pH is broadly stable across years (7.6–8.0 typical range). The 2014 outlier value (13.94) in the pivot is physically impossible and represents a data quality issue requiring CPCB follow-up.
- Nitrate peaked in 2016 (3.78 mg/L avg) and reached its lowest in 2023 (1.15 mg/L) — a positive trend. However, the grand average of 16.07 mg/L across all years still exceeds WHO limits, indicating systemic agricultural runoff challenge.

10. Dashboard Design

The project dashboard was implemented entirely in Google Sheets using pivot tables, calculated columns, and native charting tools. A dedicated 'Dashboard' sheet was created with gridlines disabled for a clean, professional layout.

10.1 Dashboard Objective

To provide a single-screen, interactive executive view enabling users to filter by Year, State, Monitoring Location, and Pollution Category — and instantly see how all charts update in response — supporting real-time question answering during presentations and policy discussions.

10.2 Dashboard Layout

Zone	Component	Chart Type	Data Source
Top Banner	Title + Key Insights text block	Static text	Written insights
Top Left	Year-wise Pollution Trend	Line Chart	Pivot 1
Top Right	State-wise Pollution Comparison	Horizontal Bar	Pivot 2
Middle Left	Pollution Count by State	Bar Chart	Pivot 3
Middle Right	Pollution Category Distribution	Donut Chart	Pivot 8

Bottom Left	State-wise Trend (Top States)	Multi-line Chart	Pivot 7
Bottom Right	Top Pollution Hotspots	Horizontal Bar	Pivot 4
Additional	Geographic Bubble Map	Geo Chart	Location + Count data
Side Panel	4 Interactive Slicers	Slicers	All pivots connected

10.3 Interactive Slicers

Four slicers were inserted (Insert → Slicer) and connected to all pivots via Report Connections:

- Year Slicer — filter all charts to specific years
- State Name Slicer — isolate analysis to selected states
- Monitoring Location Slicer — drill down to individual stations
- Pollution Category Slicer — show only Good / Moderate / Polluted stations

10.4 Key Insights Panel (from Dashboard)

- Pollution is concentrated in a small number of monitoring locations; targeted clean-up actions at hotspot stations will be more effective than broad interventions.
- Certain states consistently show higher pollution levels and should be prioritized for stricter monitoring and pollution control policies.
- Although the majority of records fall under the 'Good' category, a significant moderate and polluted share indicates the need for continuous monitoring.
- State-wise trends reveal different pollution behaviors, suggesting that region-specific strategies are more effective than uniform national approaches.
- Increasing monitoring frequency and focusing on high-risk seasons and locations can help reduce long-term pollution levels

11. Insights Summary

. The following 10 insights are derived directly from the actual pivot tables, charts, and data values confirmed from the analysis. Each is written in decision-ready language.

#1	**Insight** National BOD has declined 70% from 10.16 mg/L (2013 peak) to 3.05 mg/L (2023), indicating genuine overall improvement.	**Implication** Improvement is real but incomplete — 2023 BOD still sits at the Moderate-Polluted boundary. Policies must sustain the improvement trajectory to breach the 2 mg/L 'Good' threshold.

#2	**Insight** COVID-19 lockdowns (2020–21) produced the best water quality on record: BOD 2.38–2.71 mg/L and DO 7.53 mg/L — driven by reduced industrial and transport activity.	**Implication** Industrial discharge is a confirmed primary driver of river pollution. Targeted industrial regulation can replicate a significant portion of the lockdown-era water quality improvement.
#3	**Insight** 68.8% of records are 'Good' but 31.2% (3,660 station-years) are Moderate or Polluted — representing persistent risks to communities in river corridors.	**Implication** The majority 'Good' headline masks a significant at-risk population. Continuous monitoring of the 31.2% non-Good stations is as important as improving the worst ones.
#4	**Insight** Satluj B/C (Punjab) has average BOD exceeding 175 mg/L — 38.8 times the national average and nearly 60 times the 'Good' threshold.	**Implication** This is an acute ecological emergency, not gradual degradation. Emergency industrial audit and discharge stop orders at this station should be the #1 environmental priority action.
#5	**Insight** Maharashtra leads in pollution count (180 flagged records) but Haryana leads in BOD intensity (8.02 mg/L avg). These require different policy responses.	**Implication** Count-based and intensity-based rankings diverge — Maharashtra needs broad networkwide STP upgrades; Haryana needs targeted BOD reduction at its most severe stations.
#6	**Insight** Uttarakhand recorded 29.22 mg/L BOD in 2014 — 6.5x the national average — before recovering sharply. Delhi spiked to 16.16 mg/L in 2018.	**Implication** Episodic acute pollution events can be more dangerous than chronic moderate pollution. Realtime monitoring systems are needed to detect and respond to such spikes within days, not years.
#7	**Insight**	**Implication**
	Kerala has the lowest average DO of any major state (5.96 mg/L) despite moderate BOD (5.55 mg/L) — an unusual pattern not explained by organic pollution alone.	Single-parameter BOD analysis misses this risk. Kerala's rivers need specific investigation for chemical oxygen demand, algal growth, or water hyacinth — the cause determines the intervention.

#8	**Insight** 3 of the top 6 hotspot stations by BOD are in Maharashtra (River Wardha, Panchaganga, Mula-Mutha), explaining why it tops the pollution count ranking.	**Implication** Maharashtra's dual appearance in both count and hotspot rankings signals systemic industrial discharge problems — particularly in the Vidarbha and Pune river corridors.
#9	**Insight** Nitrate grand average of 16.07 mg/L across all station-years exceeds WHO drinking water guideline of 10 mg/L.	**Implication** Agricultural runoff is a nationwide systemic challenge independent of the industrial BOD problem. Nitrate management through precision agriculture and buffer zones must be part of any comprehensive river health strategy.
#10	**Insight** Rd_Bangalore, Tripura, and Puducherry have zero or near-zero pollution flags — the cleanest river water in the dataset.	**Implication** These regions serve as national benchmarks. Understanding their land use, industrial profile, and regulatory environment can provide bestpractice models for polluted states.

12. Recommendations

Recommendation 1: Emergency Action at Satluj B/C and Top 5 Hotspots

Mapped Insight	Insight #4 — Satluj BOD = 175+ mg/L, 38.8x national average
Action	Deploy pollution investigation teams immediately; audit all industrial discharge within 5km upstream of Satluj B/C; issue temporary closure notices to non-compliant industries; install continuous BOD sensors at all top-5 hotspot stations
States	Punjab (Satluj), Maharashtra (Wardha, Panchaganga, Mula-Mutha)
Impact	Reducing top-5 hotspots from >100 mg/L to <10 mg/L would meaningfully shift national averages and protect downstream drinking water
Feasibility	High — geographically targeted; CPCB has legal authority to issue closure orders

Recommendation 2: State-Specific Programs for Maharashtra, Haryana, MP

Mapped Insight	Insight #5 — Maharashtra: 180 flagged records; Haryana: 8.02 mg/L avg BOD
Action	Maharashtra: Expand STP capacity in Wardha, Pune, and Nasik river corridors; enforce industrial ETP compliance. Haryana: Intensive BOD reduction at the specific stations driving its high state average
Impact	Maharashtra 50% count reduction = 90 fewer flagged station-years annually. Haryana BOD reduction to national avg = state moves from #1 worst to mid-range
Feasibility	Medium — multi-year investment; Jal Jeevan Mission funding available

Recommendation 3: Real-Time Event-Spike Early Warning System

Mapped Insight	Insight #6 — Uttarakhand 2014 spike (29 mg/L) and Delhi 2018 spike (16 mg/L) not detected in time
Action	Install automated BOD sensors at all Class A monitoring stations in top-10 polluted states. Alert threshold: 2x the 3-year rolling average for that station. Connect to state PCB dashboards for same-day response
Impact	Response time reduced from months (annual data) to hours; prevents contamination from reaching water intakes
Feasibility	Medium — technology available; requires procurement and maintenance budget

Recommendation 4: Composite River Health Index (CRHI)

Mapped Insight	Insight #7 — Kerala's low DO with moderate BOD shows singleparameter BOD flag is insufficient

Action	Develop CRHI combining BOD, DO, Fecal Coliforms, Nitrate, and pH into a single weighted score. Publish monthly CRHI rankings for all states. Replace BOD-only Pollution Flag with CRHI as national standard
Impact	More accurate risk classification; reveals hidden risks like Kerala's DO crisis; improves policy targeting
Feasibility	High — purely analytical; no hardware investment required

Recommendation 5: Industrial Regulation Based on COVID Natural Experiment

Mapped Insight	Insight #2 — BOD at lowest levels during COVID lockdowns
Action	Commission sector-wise industrial discharge analysis using 2020–2021 as the baseline. Identify which industries drove the BOD improvement. Enact targeted discharge regulations for these sectors
Impact	Even 50% of lockdown-era improvement sustained = ~600–700 Moderate/Polluted records shifting to 'Good' annually
Feasibility	High — uses existing data; requires inter-ministerial coordination but no new infrastructure

13. Impact Estimation

Impact Area	Intervention	Estimated Impact	Logic
Cost Saving	Emergency remediation at top 5 hotspots	₹80–250 Cr/year saved in downstream treatment	High BOD requires 3–5x more water treatment; hotspot BOD is 30–38x above normal
Efficiency	Targeted hotspot vs broad state programs	40–60% better cost-perunit pollution reduced	Top 17 stations (<1% of locations) drive disproportionate pollution load

Service	STP upgrades in Maharashtra and MP reducing count by 50%	~780 fewer flagged stationyears/year; safer water for river corridor communities	Maharashtra alone has 180 flags; Karnataka has 146
Risk Reduction	Real-time BOD sensors	Response time: months → hours for acute events	2013 and 2018 spikes lasted years before appearing in annual CPCB reports
Environmental	Industrial regulation replicating 50% of COVID improvement	600–700 records shift from Moderate/Polluted to Good annually	COVID period reduced national avg BOD from 4.7 to 2.4 mg/L

14. Limitations

14.1 Data Issues

- Year 2015 completely absent — unexplained gap disrupts longitudinal trend continuity between 2014 and 2016.
- pH year 2014 value (13.94 in pivot) is physically impossible (pH scale is 0–14); data entry errors exist in early-year records.
- States with denser monitoring networks (Maharashtra, Karnataka) show inflated absolute pollution counts — count rankings favour states with more stations, not necessarily the most polluted environments.
- Fecal Streptococci data dropped entirely — this leaves a gap in pathogen profiling.
- Annual min/max ranges do not reveal seasonal patterns or exact event timing

14.2 Assumption Risks

- BOD threshold of 3 mg/L for Pollution_Flag is based on CPCB Class C standards. For drinking water, the threshold would be stricter (<2 mg/L), which would reclassify many 'Moderate' records as 'Polluted'.
- Median imputation assumes missing values follow the same distribution as present values — may not hold if equipment failures correlate with high-pollution events.
- The COVID BOD improvement interpretation is correlational, not causal — reduced sampling frequency during lockdowns could partially explain the apparent improvement.

14.3 What Cannot Be Concluded

- Specific pollution sources cannot be identified — the data shows contamination levels but cannot attribute them to specific industries or municipalities.
- Downstream human health impacts cannot be directly quantified without epidemiological data.
- Heavy metals, pharmaceutical residues, and microplastics are not in this dataset — major gaps for India's industrial river systems.
- Improved water quality cannot be distinguished from reduced monitoring effort in years with fewer records.

15. Future Scope

15.1 Additional Analysis Possible

- Time-series forecasting: ARIMA or linear regression on Year-wise BOD trend to project 2024–2030 national averages.
- Cluster analysis: Group monitoring stations by multi-parameter similarity to create objective pollution typologies.
- Composite River Health Index: Combine BOD, DO, Fecal Coliforms, Nitrate, and pH into a single weighted score and re-rank all states.
- Seasonal pattern analysis: If monthly data obtained, map seasonal pollution cycles (expected higher BOD in summer low-flow, lower in monsoon high-flow/dilution).

15.2 New Data Needed

- Monthly/quarterly sampling data — enables seasonal analysis and event detection.
- Industrial discharge permits database with GPS coordinates — to link discharge points to downstream monitoring readings.
- Heavy metals and pharmaceutical residue measurements — for complete river health profiling.
- River flow rate data (m^3/s) — to normalise pollution concentrations for volume effects.
- Population and sewage generation estimates per station catchment area.
- Continuous IoT-enabled BOD monitoring to replace annual min/max sampling.

16. Conclusion

This project has successfully demonstrated how a structured data analytics pipeline can transform CPCB river water quality data into actionable environmental intelligence. From raw multi-parameter measurements to an interactive dashboard with 8 pivot tables and 4 slicers, the analysis delivers findings that are both statistically grounded and immediately policy-relevant.

The analysis of 11,718 monitoring records across Indian states from 2012 to 2023 reveals a nuanced picture: nationally, river water quality is improving — BOD has fallen 70% from its 2013 peak — but this headline conceals deep geographic inequality. A small number of extreme hotspot stations (led by Satluj B/C in Punjab with BOD exceeding 175 mg/L) represent acute emergencies, while 31.2% of all monitoring events remain outside 'Good' quality thresholds.

The COVID-19 natural experiment provides the study's most powerful policy insight: when industrial and transport activity dropped in 2020–2021, national average BOD fell to its lowest ever recorded level. This confirms that industrial discharge regulation is a high-leverage intervention — more targeted and cost-effective than broad national campaigns.

The geographic map, state trend lines, and hotspot analysis together tell a clear story: pollution in India is not a diffuse national problem but a concentrated, localised challenge solvable through targeted, state-specific, station-level action backed by data.

Value delivered: 11,718 raw records → 8 pivot tables → 7 data visualisations → 10 verified insights → 5 targeted recommendations → 1 interactive dashboard. The analysis identifies that less than 1% of monitoring stations account for a disproportionate share of India's river pollution — making targeted, cost-effective intervention not just possible, but clearly defined.

17. Appendix

Appendix A: Complete Data Dictionary

Column	Parameter	Unit	Acceptable Range	Imputed Value	Action
A	Country	—	India	—	No changes
B	Year	—	2012–2023	—	Renamed from verbose format

C	State Name	—	All states	—	No changes
D	Station Code	—	Numeric	—	Blank rows removed
E	Monitoring Location	—	River station name	—	No changes
F	Min Temperature	°C	10–30	19	Imputed
G	Max Temperature	°C	10–30	29	Imputed
H	Min DO	mg/L	>4 (aquatic life)	6	Imputed
I	Max DO	mg/L	>6 (drinking)	8	Imputed
J	Min pH	pH	6.5–8.5	7.2	Imputed
K	Max pH	pH	6.5–8.5	8.2	Imputed
L	Min Conductivity	µmho/cm	<500	206	Imputed
M	Max Conductivity	µmho/cm	<500	476	Imputed
N	Min BOD	mg/L	<2 (good)	1.3	Imputed
O	Max BOD	mg/L	<3 (acceptable)	2.8	Imputed
P	Min Nitrate	mg/L	<10	0.3	Imputed
Q	Max Nitrate	mg/L	<10	1.34	Imputed
R	Min Fecal Coliforms	MPN/100ml	<100 bathing	45	Imputed

S	Max Fecal Coliforms	MPN/100ml	<100 bathing	400	Imputed
T	Min Total Coliforms	MPN/100ml	<10 drinking	225	Imputed
U	Max Total Coliforms	MPN/100ml	<10 drinking	1600	Imputed + renamed
V	Min Fecal Streptococci	MPN/100ml	<100	—	REMOVED
W	Max Fecal Streptococci	MPN/100ml	<100	—	REMOVED

Metric	Value	Source
National Grand Average BOD	4.51 mg/L	Pivot 1 Grand Total
National Grand Average DO	6.85 mg/L	Pivot 1 Grand Total
Highest year BOD (national avg)	10.16 mg/L — Year 2013	Pivot 1
Lowest year BOD (national avg)	2.38 mg/L — Year 2021	Pivot 1
Highest state avg BOD	Haryana — 8.02 mg/L	Pivot 2
Lowest state avg BOD	Rd_Bangalore — 1.55 mg/L	Pivot 2
Lowest state avg DO	Kerala — 5.96 mg/L	Pivot 2

Appendix B: Engineered Columns with Formulas

Column	Formula	Description
DO_Avg	=(H2+I2)/2	Average DO — used in all trend/comparison pivots

BOD_Avg	=(N2+O2)/2	PRIMARY KPI — average BOD, basis of all pollution analysis
pH_Avg	=(J2+K2)/2	Average pH
Nitrate_Avg	=(P2+Q2)/2	Average nitrate
FecalColiform_Avg	=(R2+S2)/2	Average Fecal Coliform
TotalColiform_Avg	=(T2+U2)/2	Average Total Coliform
Pollution_Flag	=IF(BOD_Avg>3,1,0)	1=Polluted (BOD>3 mg/L), 0=Acceptable
Pollution_Category	=IF(BOD_Avg<=2,"Good",IF(BOD_Avg<=3,"Moderate","Polluted"))	Three-tier quality classification
Year_Clean	=VALUE(RIGHT(B2,4))	Extracts 4-digit year from full year string
Water Quality Index	Composite formula	Overall WQI combining multiple parameters

Appendix C: Summary Statistics from Actual Pivot Tables

Metric	Value	Source
Most flagged state (count)	Maharashtra — 180 records	Pivot 3
Highest BOD hotspot	Satluj B/C — ~175+ mg/L	Pivot 4
Good category	8,058 records (68.8%)	Pivot 8
Moderate category	2,097 records (17.9%)	Pivot 8

Polluted category	1,563 records (13.3%)	Pivot 8
Total records analysed	11,718	Grand Total

Appendix D: References & Data Sources

- Primary Dataset: NITI Aayog NDAP — Water Quality of Indian Rivers (CPCB) — https://ndap.niti.gov.in/dataset/7078
- Working Dataset (Google Sheets): https://docs.google.com/spreadsheets/d/1bXgl9kC0ErgKqnPbxWfZOvLowm2t0_dRE8MDKbGYk/edit
- Central Pollution Control Board: https://cpcb.nic.in — National Water Quality Monitoring Programme
- WHO Guidelines for Drinking-water Quality, 4th Edition (2011)
- Bureau of Indian Standards IS 10500:2012 — Drinking Water Specifications
- Namami Gange Programme — National Mission for Clean Ganga: https://nmcg.nic.in

18. Contribution Matrix

Team Member	Dataset & Sourcing	Cleaning	KPI & Analysis	Dashboard	Report Writing	PPT	Overall Role
Adamya Tiwari	X		X	X			Project Lead

Vishuti Jamwal	X	X				X	PPT & Quality Lead
Kartik Yadav	X		X	X			Dashboard Lead
Bhavya Punj	X	X				X	Strategy Lead
Om Chimurkar	X	X		X			Data Lead
Jigyasu Kalyan	X		X		X		Analysis Lead

— End of Report —

Newton School of Technology | Data Visualisation and Analytics | 2nd Year, 4th Semester