Indian Statistical Institute, Kolkata

Statistical Methods-III Project

Bachelors of Statistics (Honours), 2023-26

A Brief Statistical Analysis of Ganges River Water Quality Data

Srijan Bhowmick

BS2352

Advisor/Instructor

Dr. Ayanendranath Basu

Professor, Interdisciplinary Statistical Research Unit

Monday 11th November, 2024

Contents

1 Abstract

This study presents a statistical analysis of water quality parameters from the Ganges River across four years (2018–2021). The dataset includes measurements of biochemical oxygen demand (BOD), dissolved oxygen, pH, temperature, and conductivity, collected from multiple monitoring stations. Descriptive statistics, including mean, median, standard deviation, and interquartile range (IQR), were computed to summarize the data and draw simple yet powerful observations, providing an overview of water quality trends. Analysis of variance (ANOVA) was then employed to explore the temporal variations in these parameters. The analysis revealed significant differences in pH levels across years, while no significant differences were observed for BOD and dissolved oxygen. The results suggest that pH is the only parameter showing notable variation over time, which could be attributed to environmental or anthropogenic factors. The findings highlight the need for continued monitoring and a deeper investigation into the factors influencing water quality in the Ganges River.

2 Introduction

The Ganges River, vital to millions of people in India, faces significant challenges related to water quality due to industrial, agricultural, and urban impacts. Monitoring key water quality parameters, such as BOD, pH, dissolved oxygen, nitrates, and conductivity, is essential for assessing the health of the river and ensuring its sustainability.

This project analyzes water quality data from the Ganges River from 2018 to 2021 using descriptive statistics and statistical methods like ANOVA and hypothesis testing. The goal is to identify trends and differences in water quality over the years, providing valuable insights for effective water management and conservation efforts.

3 Dataset Overview

The water quality data of River Ganga was sourced from the annual studies conducted by the Central Pollution Control Board(CPCB). The data contained the maximum and minimum values of various water quality indicators including biochemical oxygen demand(BOD), dissolved oxygen, pH, temperature, conductivity, fecal and total coli form. Here is a brief description including definition and impact on water quality for each of the indicators,

- Biochemical Oxygen Demand(BOD)
  - BOD measures the amount of oxygen required by microorganisms to decompose organic matter in water over a set period.
  - Higher BOD indicates more organic pollution in the water, often from sewage or agricultural runoff. This can deplete oxygen levels, harming aquatic life by reducing the oxygen available for fish and

other organisms.

- Dissolved Oxygen(DO)
  - Dissolved oxygen refers to the amount of oxygen present in water, which is essential for the survival of aquatic organisms.
  - Low DO levels can kill aquatic life, particularly fish. Low oxygen can be caused by high temperatures, excessive nutrient pollution, or high BOD. High DO levels generally indicate healthy water conditions.
- pH
  - pH measures the acidity or alkalinity of water on a scale from 0 to 14, where 7 is neutral, below 7 is acidic, and above 7 is alkaline.
  - Extremes in pH can be harmful to aquatic life. Acidic water (pH below 6) can harm fish, disrupt ecosystems, and leach toxic metals. Alkaline water (pH above 9) can also be harmful and affect species that are sensitive to pH fluctuations.
- Conductivity
  - Conductivity measures the ability of water to conduct electricity, which increases with the presence of dissolved salts, minerals, or other inorganic substances.
  - High conductivity typically indicates high levels of dissolved solids, which could be harmful to aquatic organisms. It may result from pollution, such as runoff from industrial waste, sewage, or salts from roads.
- Nitrates
  - Nitrates are nitrogen compounds commonly found in fertilizers and sewage. Elevated nitrate levels in water can indicate nutrient pollution.
  - High nitrate levels can lead to eutrophication, causing algal blooms that deplete oxygen and harm aquatic life. Nitrates can also contaminate drinking water, leading to health risks, particularly for infants (blue baby syndrome).
- Fecal and Total Coli forms
  - Fecal coli forms are bacteria found in the intestines of warm-blooded animals, including humans, whereas total coli forms are a larger group of coli form bacteria found in the environment, including soil and plants, as well as in feces.
  - Presence of either set of coli forms in water indicate potential contamination with harmful pathogens that can cause diseases like gastroenteritis.

The following table should provide a clear overview on the admissible levels of each indicator,

Indicator	**Units**	**For Aquatic Life**	**For Drinking Water**
BOD	mg/L	< 5	< 1
Dissolved Oxygen	mg/L	> 5	$5 - 8$
Conductivity	µmhos/cm	< 1500	< 1000
pH	-	6.5 - 9	6.5 - 8.5
Nitrates	mg/L	< 10	< 10
Fecal Coliforms	-	As low as possible	0 fecal coliforms per 100 mL
Total Coliforms	-	As low as possible	0 total coliforms per 100 mL

Table 1: Admissible Levels for Water Quality Indicators

For each record there were entries pertaining to the name of the monitoring station, station code (which is uniquely assigned) and the state in which the station lies. The River Ganga flows through 5 states, Uttarakhand, Uttar Pradesh, Bihar, Jharkhand and West Bengal, so there were only 5 unique values corresponding to the STATE NAME column. There were some sparse records with the STATE NAME entry as "Interstate", but they were not relevant in the final analysis.

4 Methodology

4.1 Data Processing

In the annual studies, the original water quality data was in PDF format with non-selectable text and contained tables possibly created in MS Word. Hence, to make it compatible for R, I used online OCR tools to convert the data from PDF to Word, followed by conversion into Excel files using online tools. A thorough semi-automated check was performed to ensure correctness of data throughout the two-fold conversion process.

Some entries in the original PDF and Excel files contained only a hyphen indicating non-availability(NA) and some contained the abbreviation "BDL" which means "Below Detectable Limit", so I naturally replaced all BDL occurrences by 0 and I also removed hyphens to ensure that R registered them as NA values.

After conversion to tibbles in R (smaller and compact version of data frames in R), I created new tibbles which contained only the average values of the water quality indicators to reduce the complexity of the dataset and for further analysis using Descriptive Statistics.Throughout the subsequent analysis, I only used the dataset with the averaged values.This helped reduce the number of columns from 19 to 12.

Before performing ANOVA for the years 2018 to 2021, I decided to perform ANOVA on only the records of those monitoring stations, which collected data in every year between 2018 and 2021(both included) to avoid inaccurate inferences. There were 95, 96, 99 and 99 unique records for the respective years from 2018 to 2021, out of which there were 91 common station codes, i.e. 91 monitoring stations common across these 4 years. Hence in the final combined dataset of these 4 years, there were 91*4=364 rows and 12 columns.

While performing ANOVA, outliers were detected using Q-Q and box-and-whiskers plots and I paid attention to natural outliers which were not removed as they reflected actual environmental variability.

4.2 Statistical Methods

For the year 2021, I calculated descriptive statistics such as mean, median, standard deviation and interquartile range(IQR) of the original dataset and also the mean of each indicator for each state to summarize the key features of the water quality dataset. I also performed hypothesis testing using Analysis of Variance (ANOVA) to assess whether there were significant differences in the water quality indicators across the years 2018-2021. Before applying ANOVA, I tested the assumption of normality of residuals using Q-Q plots and box-and-whiskers plots. Due to violations of normality in some cases, ANOVA was not applicable for certain variables like temperature and conductivity and hence I excluded these results from the final analysis.

4.3 Data Analysis Procedure

The data was analyzed using R (version 4.3.1). For hypothesis testing, ANOVA was performed for BOD, dissolved oxygen, and pH levels. Q-Q plots and boxplots were used for assumption testing, while summary statistics helped identify trends and central tendencies in the data. The results were interpreted to understand how water quality has evolved over the years.

4.4 Limitations

The primary limitation of the dataset lies in missing values for some years and the presence of outliers that were deemed to reflect natural fluctuations. While ANOVA is a powerful tool for hypothesis testing, it may not be the most appropriate in the presence of non-normal data, which was the case for some of the variables tested. Also there was an inherent simplification of the data by taking the average value of each indicator, thus incurring potential data loss.

5 Descriptive Statistics

5.1 Overall Descriptive Statistics for Each Indicator for the Year 2021

5.1.1 Table

The following table summarizes key descriptive statistics for the water quality indicators in the dataset,

Indicator	Mean	Median	SD	Minimum	Maximum	IQR
Temperature (°C)	22.66	23.5	3.63	0.1	28.5	2.5
pH	7.74	7.8	0.36	5.95	8.35	0.35
Dissolved Oxygen (mg/L)	8.19	8.35	1.06	5.7	10.4	1.43
Conductivity (µS/cm)	612.91	316	1959.24	110	16514	78.5
BOD (mg/L)	2.35	2.15	0.89	1	4.45	1.08
Nitrates (mg/L)	1.09	0.64	2.38	0.32	17.7	0.24
Fecal Coliforms (per 100mL)	44138.71	14950	49311.64	2	271650	79950
Total Coliforms (per 100mL)	71682.53	49450	126176.10	2	800850	80375

Table 2: Summary of Descriptive Statistics for Water Quality Indicators

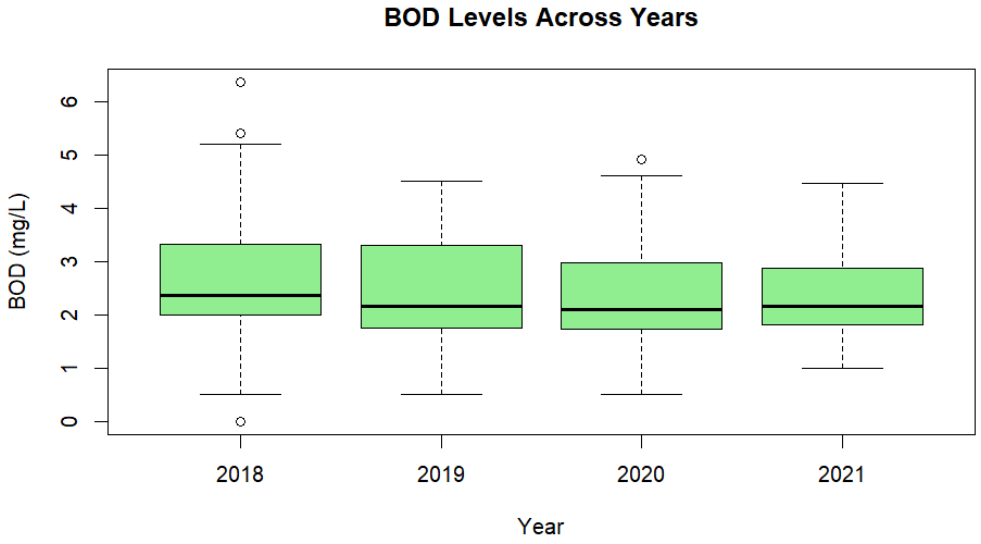
5.1.2 Key Inferences

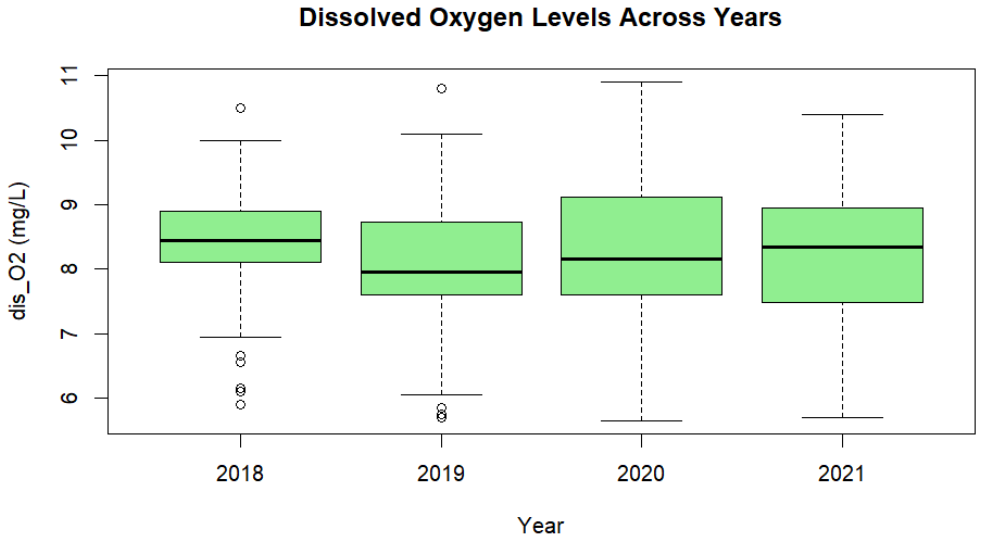
- The water temperature is moderately high, with some variability (SD = 3.63°C), suggesting seasonal or geographical differences. The minimum temperature recorded (0.1°C) indicates some extreme cold areas, possibly higher altitudes.
- The dissolved oxygen levels are relatively high, which is good for aquatic life. However, the lower end of the range (5.7 mg/L) suggests there might be some stations with lower oxygen, which could be influenced by factors like pollution or water flow.
- The pH is slightly alkaline on average, which is within the acceptable range for most aquatic environments. The low pH (5.95) in some areas could indicate acidic water, possibly due to pollution or industrial runoff.
- The wide range and high SD (1959.24 μmhos/cm) suggest significant variability in water conductivity, which can be influenced by dissolved salts, minerals, and pollution levels. The maximum value (16514 μmhos/cm) is very high, possibly indicating pollution or brackish water in certain areas.
- The BOD is relatively low, suggesting good water quality in some areas. However, higher BOD values (up to 4.45 mg/L) could indicate organic pollution in some locations.
- Nitrate levels are variable, with a high SD (2.38 mg/L) suggesting differences in pollution or runoff across different regions. The high maximum value (17.7 mg/L) could indicate excessive agricultural runoff or industrial contamination.
- Fecal coli form contamination is extremely high in some areas, with values reaching over 270,000 MPN/100 ml. This suggests serious pollution from untreated sewage or other human waste, posing significant health risks.

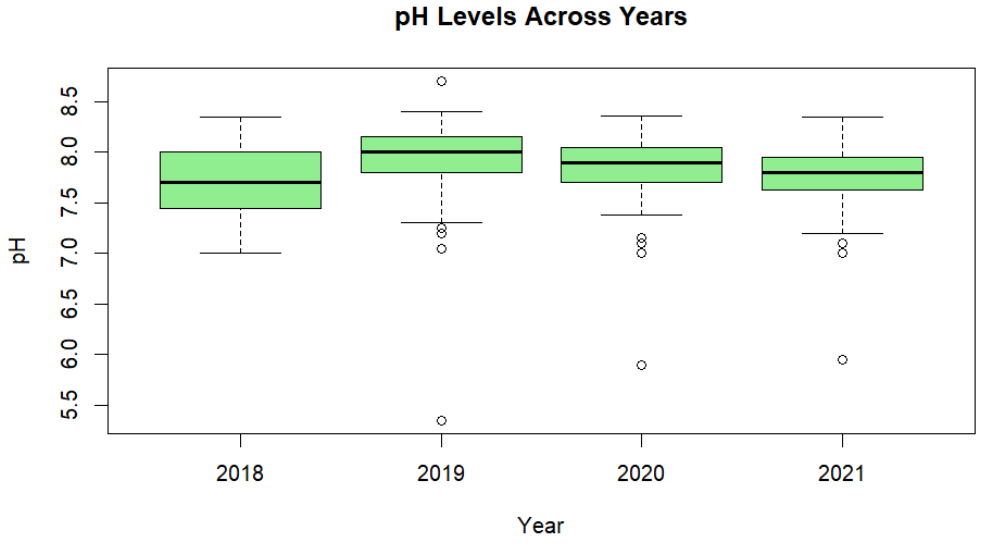
- The total coli form levels are also very high, which further supports the concerns regarding water contamination, particularly with respect to hygiene and waterborne diseases.

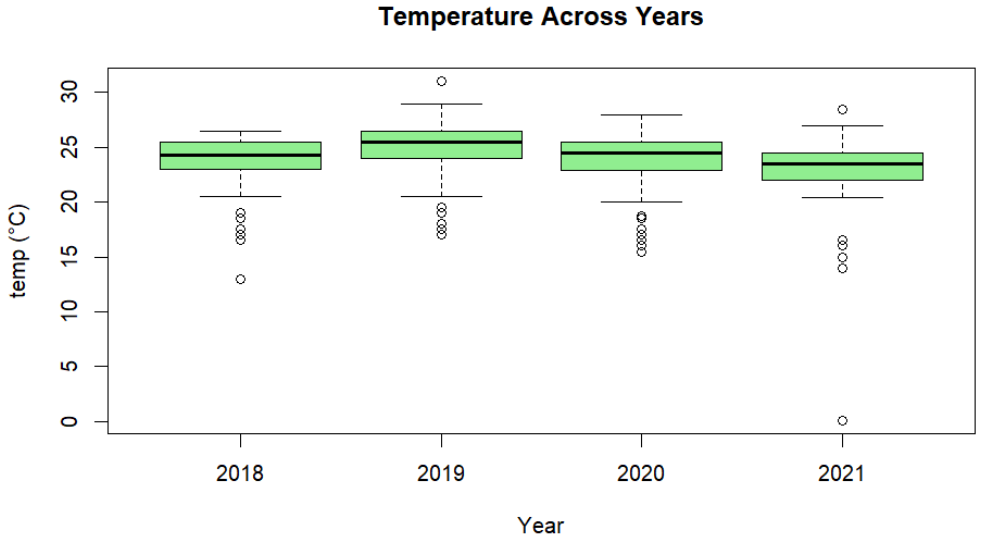
5.1.3 Plots of Indicators Across Years

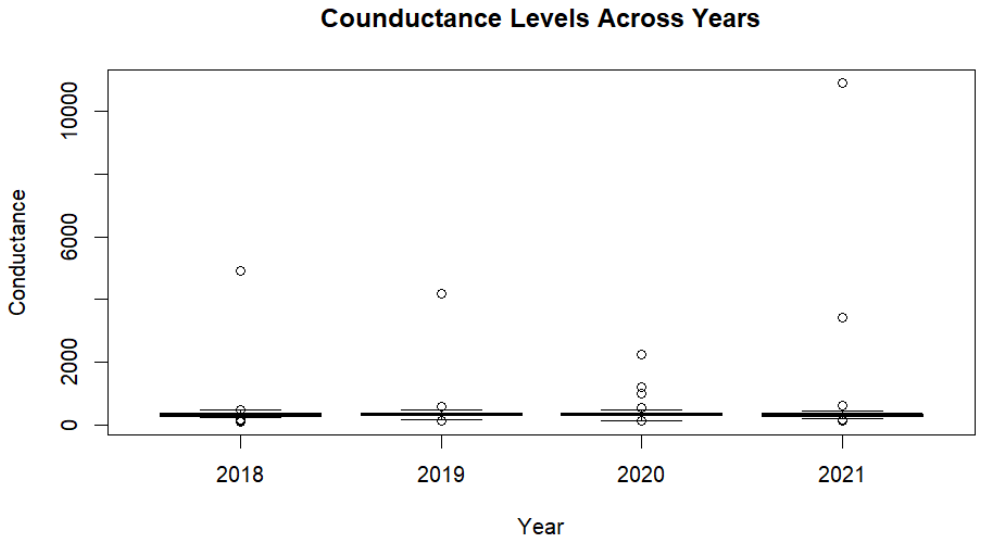
- The dataset avg_data_common was used to produced the following box-and-whiskers plots.
- These plots do not necessarily relate to the previous table because the table summarized the key descriptive statistics for 2021 whereas these plots are made across years 2018-2021 to provide a visual idea about the trends, ranges and outliers of these indicators.
- These plots are relevant for our subsequent ANOVA analysis.



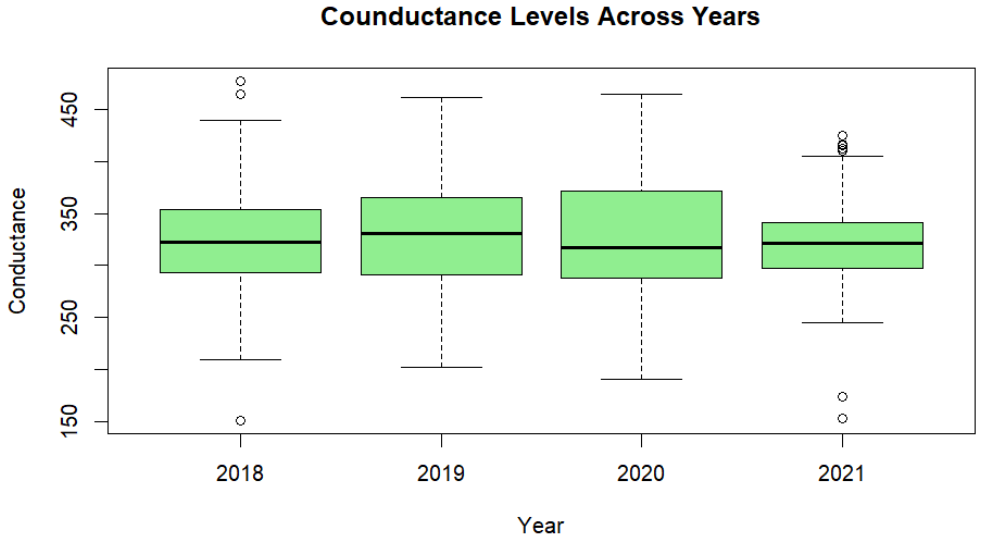



pH Levels Across Years
pH
5.5
6.0
6.5
7.0
7.5
8.0
8.5
2018
2019
2020
2021
Year


Temperature Across Years
temp (°C)
0
5
10
15
20
25
30
2018
2019
2020
2021
Year



- The following plot graphs conductance levels without outliers across years to better visualize the spread of the majority of the data across years.
- The outliers cannot be removed in order to perform ANOVA because those outliers were due to natural factors.
- Only those points were considered as outliers which did not fall into the IQR range.
- Please refer to the source code linked in the references section for a list of these outliers.



5.2 State-wise Analysis of Each Indicator

5.2.1 Tables

The following tables summarizes the state-wise averages of each indicator across 2021, (refer to above table for units)

	State Name	Temp. Mean	DO Mean	pH Mean	Conduct. Mean
1	**Uttarakhand**	16.50	9.68	7.56	209.13
2	**Uttar Pradesh**	23.81	8.44	7.84	627.14
3	**Bihar**	22.62	8.11	7.77	354.79
4	**Jharkhand**	25.12	7.06	7.59	307.00
5	**West Bengal**	26.07	6.55	7.69	1727.25

Table 3: State-wise averages for Temperature, Dissolved Oxygen, pH, and Conductivity in 2021

	State Name	BOD Mean	Nitr. Mean	Fec. Coli. Mean	Tot. Coli. Mean
1	**Uttarakhand**	1.27	0.58	825.67	53374.63
2	**Uttar Pradesh**	2.84	0.94	7250.19	34890.31
3	**Bihar**	2.30	1.49	86599.59	88128.85
4	**Jharkhand**	1.57	–	–	–
5	**West Bengal**	2.71	0.84	71742.86	135453.57

Table 4: State-wise averages for BOD, Nitrate, Fecal Coliform, and Total Coliform in 2021

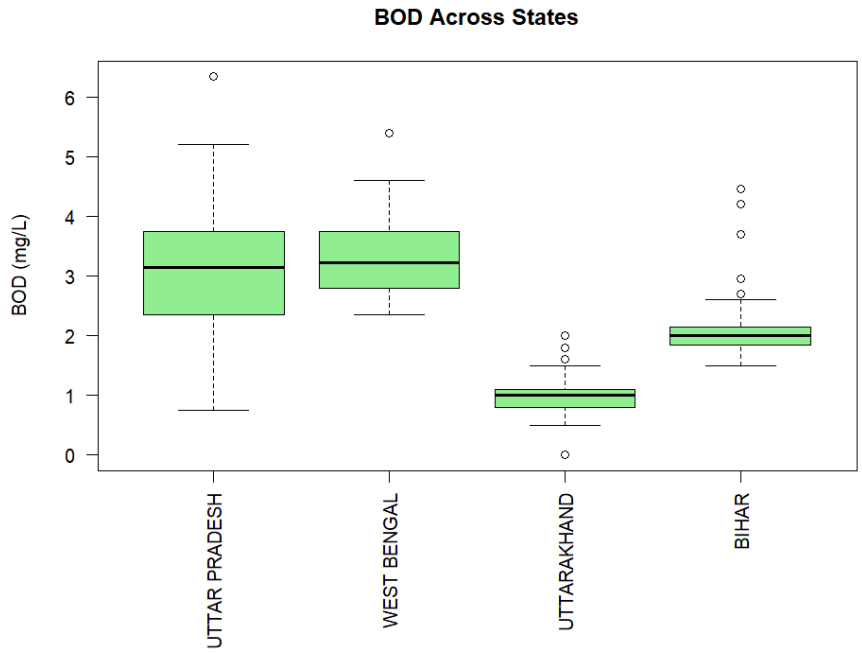
- Note that the values for the means of nitrate, fecal coli form and total coli form are missing because there are only 4 monitoring stations located in Jharkhand which failed to collect this data possibly because of lack of funding and other reasons.

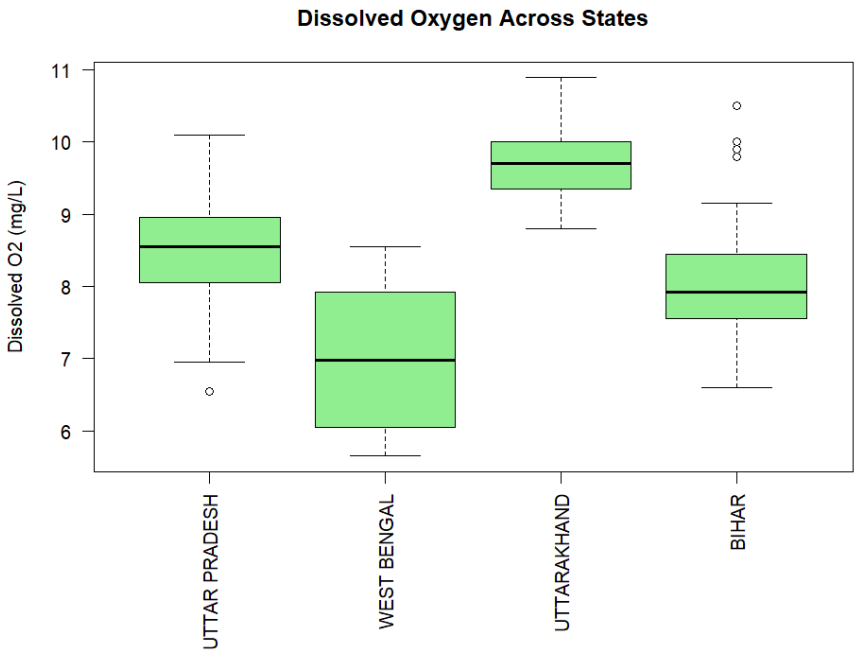
5.2.2 Key Inferences

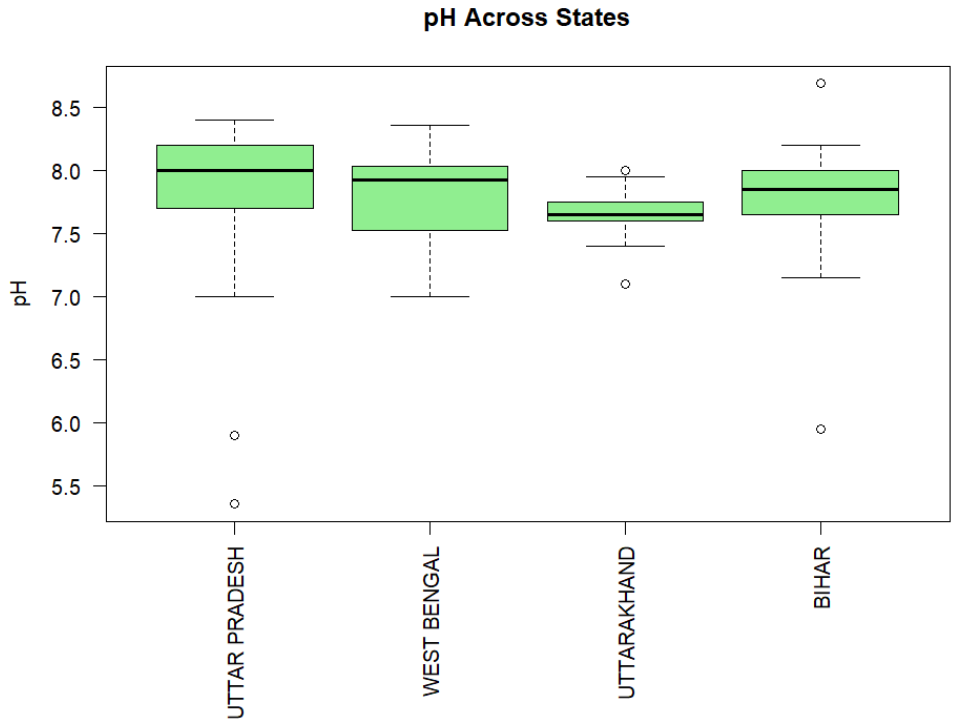
- Uttarakhand has the lowest average temperature (16.5°C) and relatively high dissolved oxygen levels (9.68 mg/L) because of its geographical location.
- Uttar Pradesh shows relatively higher BOD (2.84 mg/L) and fecal coli form (7,250 MPN/100 ml) compared to other states, indicating pollution issues.
- Bihar has a high level of fecal coli form contamination (86,600 MPN/100 ml) and a very high total coli form level (88,129 MPN/100 ml), pointing to severe water quality concerns.
- Jharkhand has missing data on nitrate and coliforms, but its average temperature (25.1°C) and pH (7.59) suggest more neutral conditions.
- West Bengal has the highest average conductivity (1,727 μmhos/cm) and total coli form (135,454 MPN/100 ml), indicating serious issues with salinity and pollution.

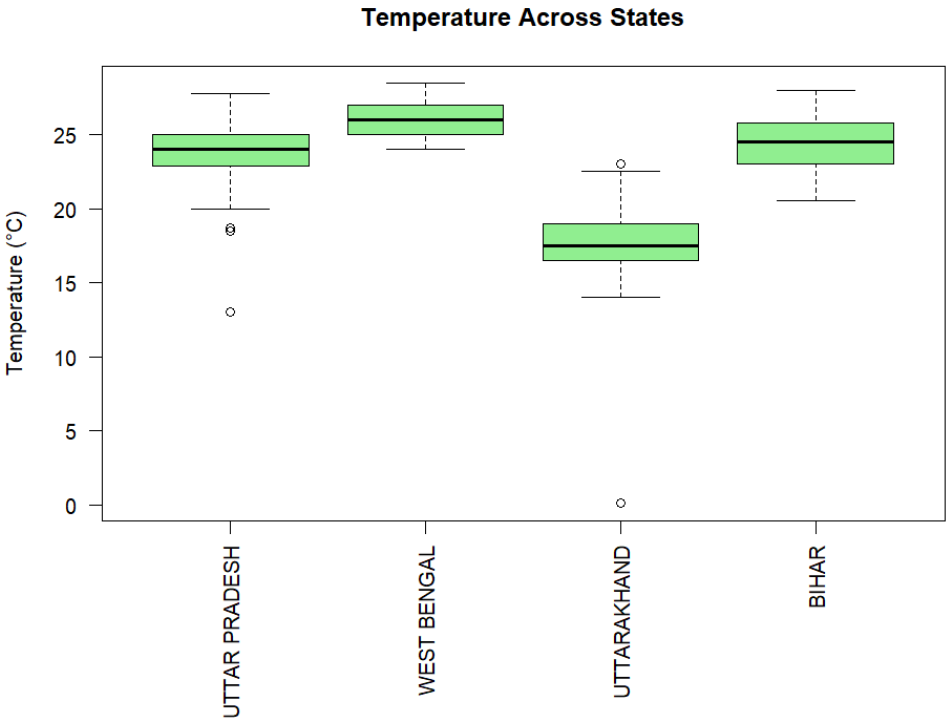
5.2.3 Plots Across States

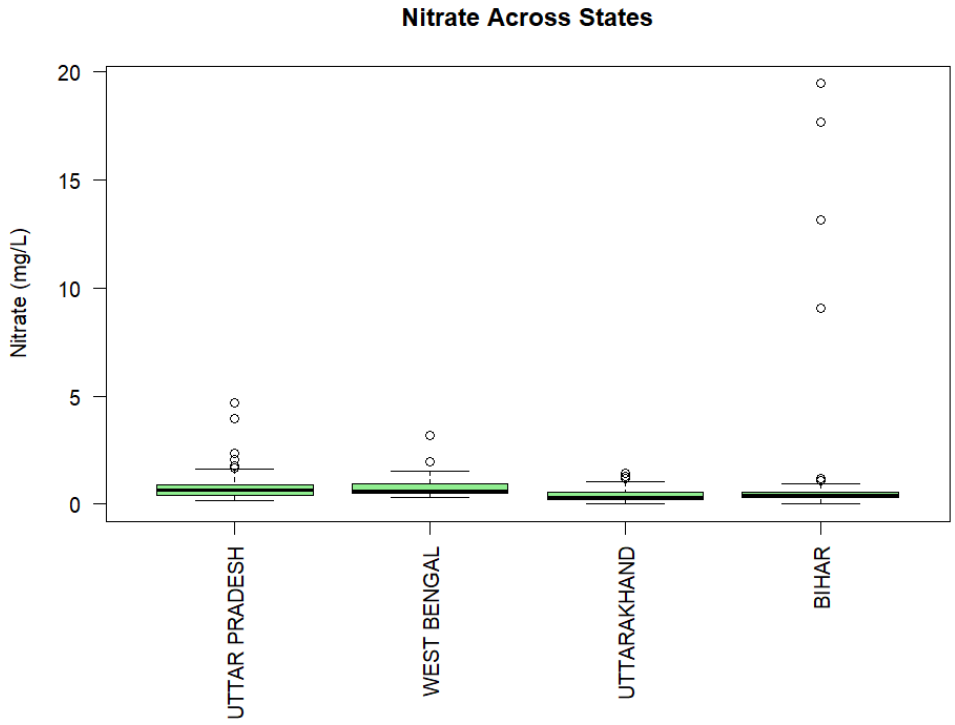
- The dataset filtered_data was used to generate the following plots of indicators across states.
- There are only 4 monitoring locations in Jharkhand. It turns out that after combining the data of common monitoring locations from years 2018 to 2021, none of the 4 above mentioned locations made the final cut. Hence I created a new dataset filtered_data to filter out records for Jharkhand to avoid a empty slot for Jharkhand in the following box plots.

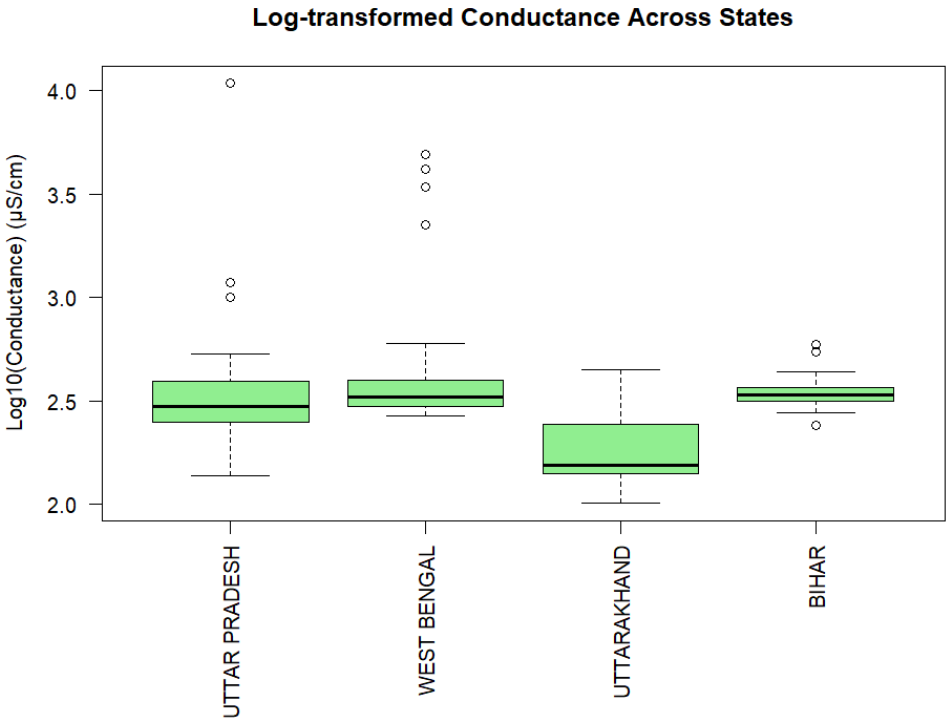


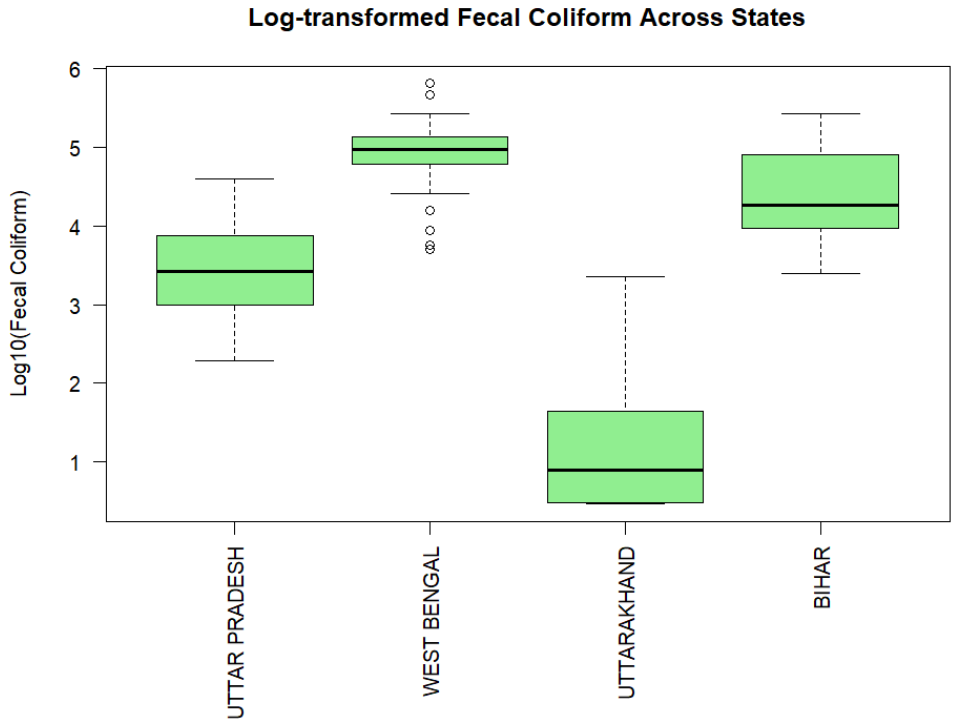


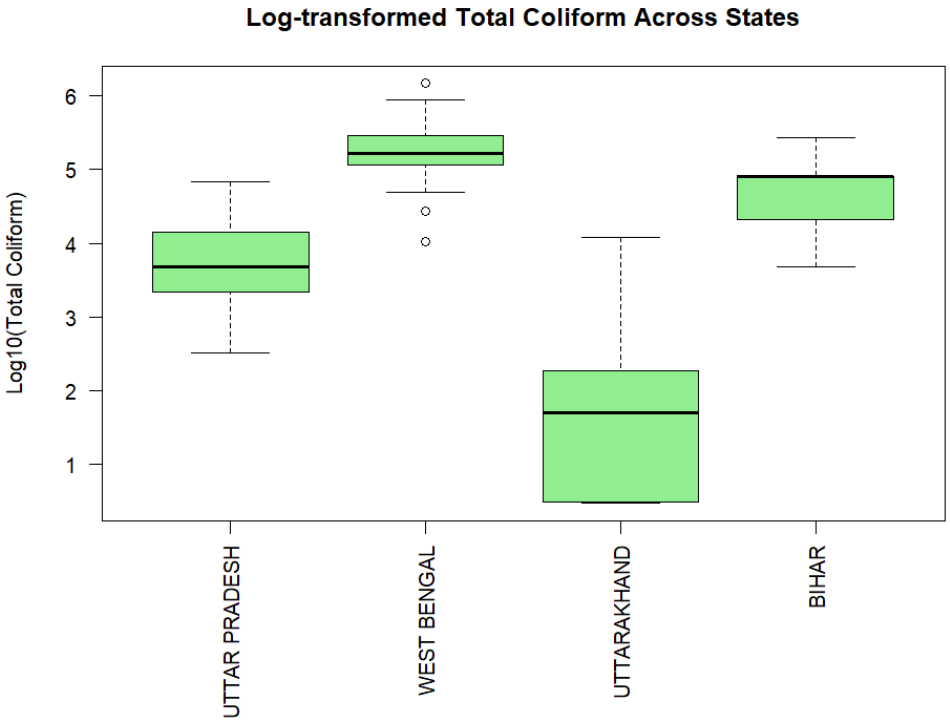

pH Across States
pH
8.5
8.0
7.5
7.0
6.5
6.0
5.5
UTTAR PRADESH
WEST BENGAL
UTTARAKHAND
BIHAR


Temperature Across States
Temperature (°C)
25
20
15
10
5
0
UTTAR PRADESH
WEST BENGAL
UTTARAKHAND
BIHAR


Nitrate Across States
Nitrate (mg/L)
20
15
10
5
0
UTTAR PRADESH
WEST BENGAL
UTTARAKHAND
BIHAR


Log-transformed Conductance Across States
Log10(Conductance) (µS/cm)
4.0
3.5
3.0
2.5
2.0
UTTAR PRADESH
WEST BENGAL
UTTARAKHAND
BIHAR


Log-transformed Fecal Coliform Across States
Log10(Fecal Coliform)
6
5
4
3
2
1
UTTAR PRADESH
WEST BENGAL
UTTARAKHAND
BIHAR


Log-transformed Total Coliform Across States
Log10(Total Coliform)
6
5
4
3
2
1
UTTAR PRADESH
WEST BENGAL
UTTARAKHAND
BIHAR

6 ANOVA Results Across Years

6.1 Overview

I conducted hypothesis testing by employing ANOVA to investigate whether there were significant differences in the mean values of various water quality indicators across the years (2018-2021) in the combined dataset. I chose ANOVA to determine if changes over time were statistically significant, particularly for key parameters like pH, BOD (Biochemical Oxygen Demand), and Dissolved Oxygen. Q-Q plots were also visually inspected to ensure that the assumption of normality holds for pH, BOD and Dissolved Oxygen. Hence, in each case, my null hypothesis (H_0) is that there is no significant difference in the mean values of the indicator across the years. My alternate hypothesis (H_1) is the negation of the null hypothesis i.e., there is a significant difference in the mean values of the indicator across at least one pair of years.

6.2 Indicator-Wise Results

6.2.1 Bio-Chemical Oxygen Demand

The following table summarizes the ANOVA results for BOD,

SoV	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Between Years	3	4.36	1.45	1.41	0.2395
Residuals	359	370.34	1.03		
Total	362	374.7			

Table 5: ANOVA Summary for BOD Levels Across Years 2018-2021

- Note that the total degrees of freedom is 362 instead of 364-1=363 since there was 1 missing value in the dataset.
- For BOD, the ANOVA test showed no statistically significant difference between the years analyzed ($p = 0.2395 > 0.05$). This suggests that BOD levels have remained relatively consistent, indicating stable organic matter levels over time without substantial annual variations.

6.2.2 Dissolved Oxygen

The following table summarizes the ANOVA results for Dissolved Oxygen,

SoV	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Between Years	3	4.08	1.36	1.33	0.2659
Residuals	360	369.28	1.03		
Total	363	373.36			

Table 6: ANOVA Summary for Dissolved Oxygen Levels Across Years 2018-2021

- There was no missing value in the dataset hence the total degrees of freedom is 363.
- The ANOVA test for dissolved oxygen levels also indicated no significant difference across years ($p = 0.2659 > 0.05$). Thus, dissolved oxygen in the Ganges has not shown marked year-to-year variations, suggesting stability in oxygenation levels, which implies steady river health in terms of oxygen content.

6.2.3 pH

The following table summarizes the ANOVA results for pH values,

SoV	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Between Years	3	2.14	0.71	5.62	0.0009
Residuals	359	45.52	0.13		
Total	362	47.66			

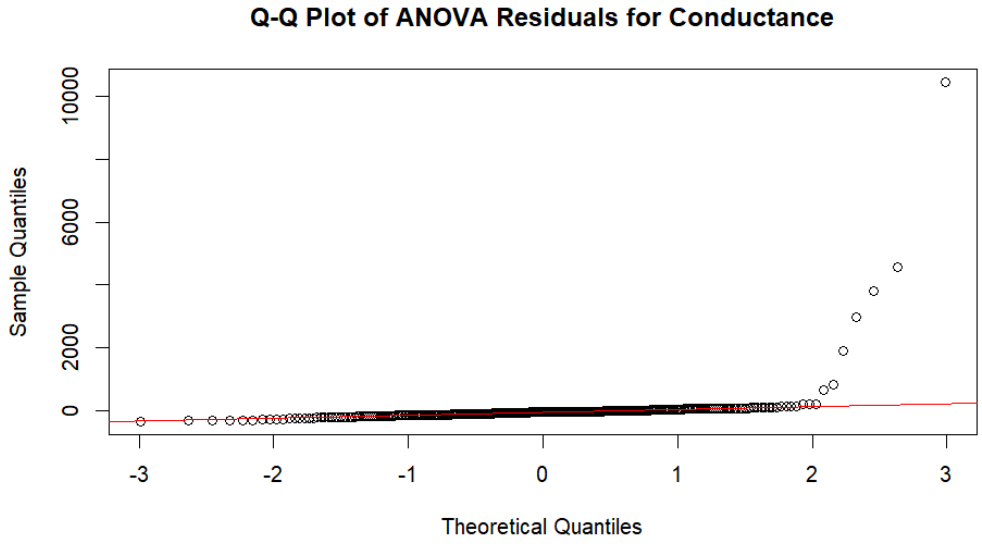
Table 7: ANOVA Summary for pH Levels Across Years 2018-2021

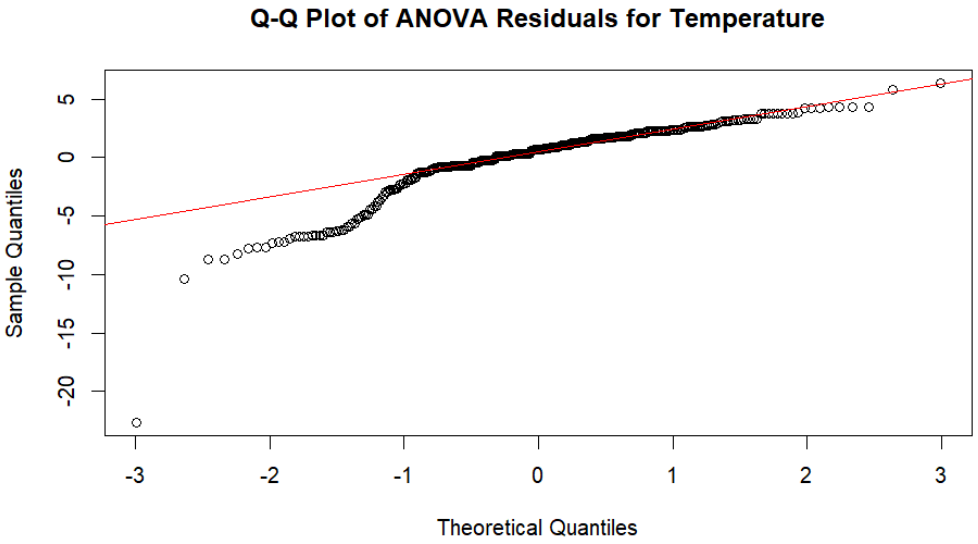
- Note that the total degrees of freedom is 362 instead of 364-1=363 since there was 1 missing value in the dataset.
- The ANOVA results for pH indicate a statistically significant difference across years ($p = 0.0009 < 0.05$). This finding suggests that pH levels in the Ganges River have fluctuated notably over time, reflecting shifts in pollution sources, treatment efforts, or natural environmental changes.
- From the box and whiskers plot of pH levels across years, it can be visually inferred that there was indeed a significant difference average pH values amongst the 4 years.

6.3 Indicators with Normality Concerns

6.3.1 Temperature and Conductance

Here are the Q-Q plots for the ANOVA residuals of conductance and temperature,





- Both temperature and conductance indicators displayed clear deviations from normality, with their Q-Q plots showing significant departures from the straight line, especially at the tails. This suggests that these indicators may be skewed or have heavy tails.
- There are simple geographical reasons for the deviations from normality. The River Ganga flows through multiple states which have significant temperature differences amongst each other, Uttarakhand being significantly colder than other. This is also clear from the box and whiskers plot of temperature across states.
- Conductance had several outliers with very high conductances, after a careful analysis of the monitoring locations of these outliers, quite a few of them were near barrages where the reduction in flow may lead to water stagnation, allowing for higher evaporation rates, which can concentrate dissolved solids and increase conductance.

6.3.2 Fecal and Total Coli forms

- Fecal and total coli form counts exhibited extreme positive skewness, with numerous outliers. These indicators had distributions that were highly right-skewed, with a long tail in the higher range of values.
- Extreme water pollution in Uttar Pradesh, Bihar and West Bengal is the chief cause for such extreme positive skewness.
- Due to the presence of many zero values and some extremely large values, a log transformation was applied to reduce the skewness and to properly visualize the data using a box and whiskers plot. (Refer to the prior plots section)

6.4 Summary

In summary, the ANOVA analysis revealed a statistically significant year-to-year difference only for pH levels, with no significant changes found for Bio-Chemical Oxygen Demand and Dissolved Oxygen. This outcome highlights a potential shift in water acidity over time, while other water quality parameters have remained relatively stable, although, further analysis of these indicators using non-parametric tests across states could lead to potentially more interesting inferences.

7 Conclusion

In this study, I analyzed various water quality indicators from the Ganges River over multiple years, with a focus on key metrics such as temperature, pH, dissolved oxygen, BOD, nitrate levels, and fecal coli forms. The analysis aimed to understand trends, assess the variation across different states and identify any significant differences across years.

Key findings include:

- State-wise Variability: Significant differences were found in the water quality indicators across different states, indicating regional variations in water quality. Conductance was quite high in certain locations in West Bengal and Uttar Pradesh and generally low in Uttarakhand. Uttarakhand had an overall better water quality than the other states. The chief source of pollution was fecal and total coli forms in Bihar, West Bengal and Uttar Pradesh which made water from River Ganga unusable. ANOVA was not applied across states due to the clear violations of the assumption of normality.

- Temporal Trends: Across the years, ANOVA results for pH showed significant difference over the years, whereas other indicators such as BOD and dissolved oxygen remained relatively stable.

- Normality Issues: Several indicators, such as fecal coli forms and total coli forms, exhibited severe positive skewness, suggesting the need for a logarithmic transformation to meet normality assumptions for statistical testing. This skewness was due to the presence of outliers and many zero values.

- Management Implications: The results highlight the importance of regular monitoring of water quality across different states and over time. States showing high levels of pollution (e.g., fecal coli forms) such as Bihar and West Bengal may benefit from more stringent water treatment practices and pollution control measures. Furthermore, the results suggest that water quality in the River Ganga may require targeted interventions near areas around particular monitoring locations, especially in states with poor water quality indicators.

8 References

[1] CPCB Website [Link]

[2] Github Repository [Link]