Water Quality Analysis

Widespread water contamination is a daunting challenge in India with 195,813 habitations in the country reported to have poor water quality, posing serious health risks to the population. Pollution in India has led to more than 2.3 million premature deaths in 2019, according to a Lancet study. Nearly 1.6 million deaths were due to air pollution, and more than half a million were caused by water pollution.

This dataset comprises water quality parameters collected from various monitoring stations across different states in India. It covers the period from 2005 to 2014 and includes information on physical, chemical, and microbiological indicators of water quality.

Description of dataset features

- **STATION CODE**: Unique identifier for each monitoring station.
- **LOCATIONS**: Geographic location of the station.
- **STATE**: State or territory where the station is located.
- **Temp**: Water temperature in degrees Celsius.
- **D.O. (mg/l)**: Dissolved oxygen concentration in milligrams per litre.
- **PH**: Water acidity or alkalinity on a scale of 0 to 14.
- **CONDUCTIVITY (µmhos/cm)**: Ability of water to conduct electric current, measured in micromhos per centimetre.
- **B.O.D. (mg/l)**: Biological oxygen demand, indicating organic matter in the water, measured in milligrams per litre.
- **NITRATENAN N+ NITRITENANN (mg/l)**: Nitrate-nitrogen, Nitrite-nitrogen and Ammonium-nitrogen concentration in milligrams per litre.
- **FECAL COLIFORM (MPN/100ml)**: Presence of faecal coliform bacteria, indicating possible faecal contamination, measured in Most Probable Number per 100 millilitres.
- **TOTAL COLIFORM (MPN/100ml)Mean**: Average Presence of total coliform bacteria, indicating potential water quality issues, measured in Most Probable Number per 100 millilitres.
- **year**: Year in which the data was collected.

Observation per variables

The provided data consists of the following number of observations for each feature:

- **STATION CODE**: 1869 observations

- **LOCATIONS**: 1807 observations
- **STATE**: 1230 observations
- **Temp**: 1899 observations
- **D.O. (mg/l)**: 1960 observations
- **PH**: 1983 observations
- **CONDUCTIVITY (µmhos/cm)**: 1966 observations
- **B.O.D. (mg/l)**: 1948 observations
- **NITRATENAN N+ NITRITENANN (mg/l)**: 1766 observations
- **FECAL COLIFORM (MPN/100ml)**: 1675 observations
- **TOTAL COLIFORM (MPN/100ml)Mean**: 1859 observations
- **year**: 1991 observations

These counts represent the number of data points available for each respective feature in the dataset.

Percentage of missing values

The given values represent the percentage of missing values for each feature in the dataset:

- **STATION CODE:** 6.13%
- **LOCATIONS:** 9.24%
- **STATE:** 38.22%
- **Temp:** 4.62%
- **D.O. (mg/l):** 1.56%
- **PH:** 0.40%
- **CONDUCTIVITY (µmhos/cm):** 1.26%
- **B.O.D. (mg/l):** 2.16%
- **NITRATENAN N+ NITRITENANN (mg/l):** 11.30%
- **FECAL COLIFORM (MPN/100ml):** 15.87%
- **TOTAL COLIFORM (MPN/100ml)Mean:** 6.63%
- **year:** 0.00%

These percentages indicate the proportion of missing data for each respective feature in the dataset.

Type of Variables

- **STATION CODE:** Categorical

- **LOCATIONS:** Categorical
- **STATE:** Categorical
- **Temp:** Continuous
- **D.O. (mg/l):** Continuous
- **CONDUCTIVITY (µmhos/cm):** Continuous
- **B.O.D. (mg/l):** Continuous
- **NITRATENAN N+ NITRITENANN (mg/l):** Continuous
- **FECAL COLIFORM (MPN/100ml):** Continuous
- **TOTAL COLIFORM (MPN/100ml)Mean:** Continuous
- **year:** Categorical

Duplicate entries

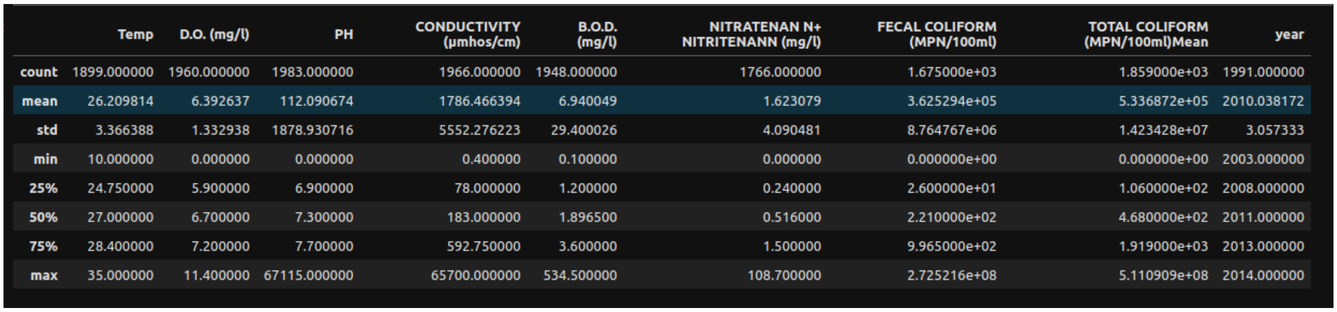
The dataset contains a single duplicate entry.

Problem Statement

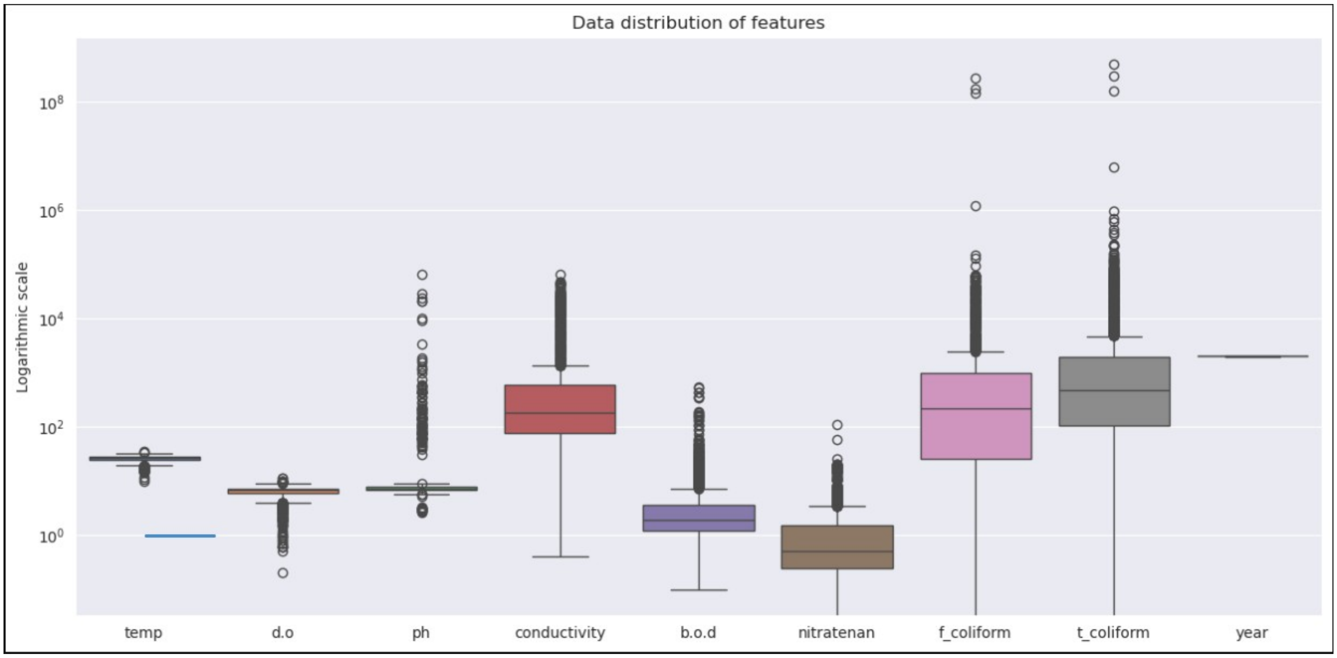
The analysis aims to explore trends and patterns in water quality, identify potential pollution sources, assess risks to human health and aquatic life, and support informed decision-making for water management and pollution control.

Data Cleaning

Handling Extreme Values



	Temp	D.O. (mg/l)	PH	CONDUCTIVITY (µmhos/cm)	B.O.D. (mg/l)	NITRATENAN N+ NITRITENANN (mg/l)	FECAL COLIFORM (MPN/100ml)	TOTAL COLIFORM (MPN/100ml)Mean	year
count	1899.000000	1960.000000	1983.000000	1966.000000	1948.000000	1766.000000	1.675000e+03	1.859000e+03	1991.000000
mean	26.209814	6.392637	112.090674	1786.466394	6.940049	1.623079	3.625294e+05	5.336872e+05	2010.038172
std	3.366388	1.332938	1878.930716	5552.276223	29.400026	4.090481	8.764767e+06	1.423428e+07	3.057333
min	10.000000	0.000000	0.000000	0.400000	0.100000	0.000000	0.000000e+00	0.000000e+00	2003.000000
25%	24.750000	5.900000	6.900000	78.000000	1.200000	0.240000	2.600000e+01	1.060000e+02	2008.000000
50%	27.000000	6.700000	7.300000	183.000000	1.896500	0.516000	2.210000e+02	4.680000e+02	2011.000000
75%	28.400000	7.200000	7.700000	592.750000	3.600000	1.500000	9.965000e+02	1.919000e+03	2013.000000
max	35.000000	11.400000	67115.000000	65700.000000	534.500000	108.700000	2.725216e+08	5.110909e+08	2014.000000

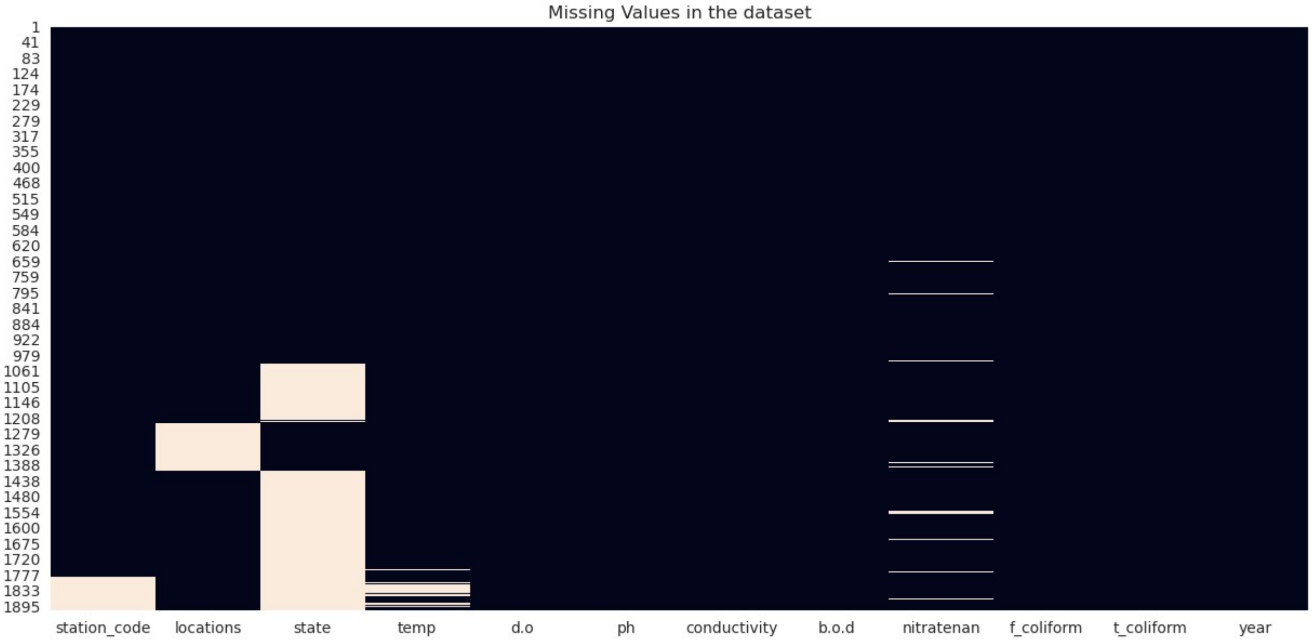


Comparing the box plot with the Statistical properties of features, some of the features contain outliers and some extreme values that will affect the quality of the analysis. Features like PH, conductivity, b.o.d, nitratenan, f_coliform and t_coliform. The maximum and minimum values for these features suggest some form of data entry error.

These outliers are treated by using the data between the 5th percentile and to 90th percentile of these features.

Handling Missing Values

Visualizing the missing data in the dataset:

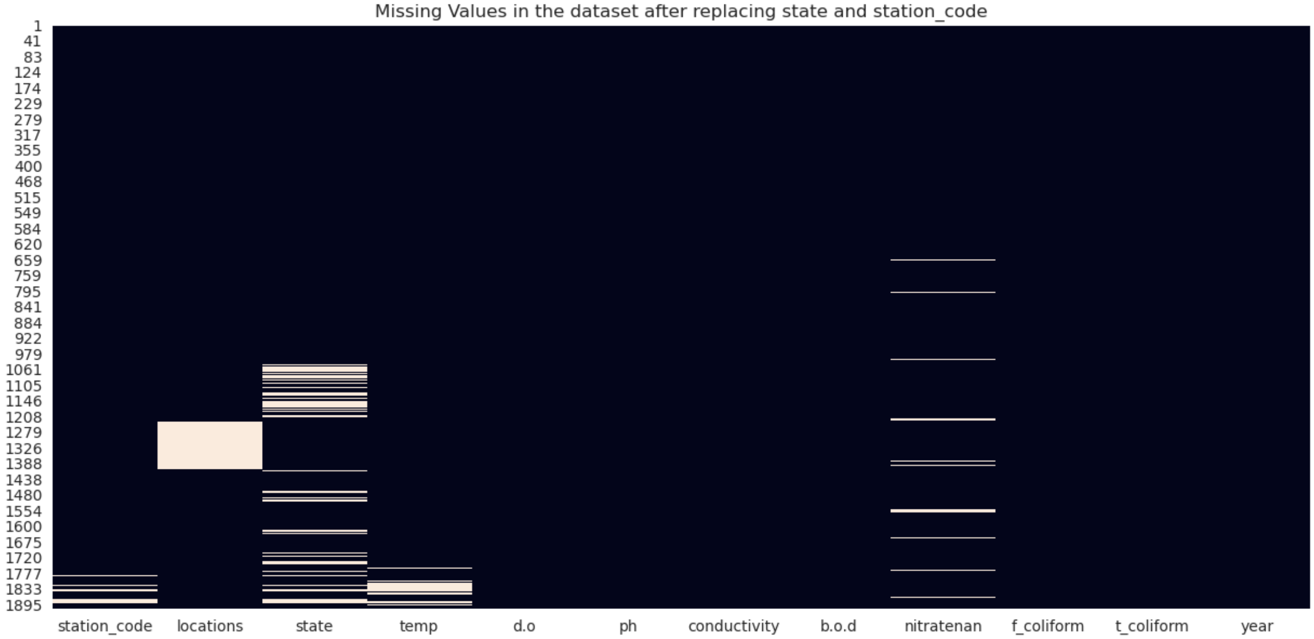


The majority of missing values from the dataset can be seen in the “state” feature. Other features like “station_code”, “locations”, “temp” and “nitratenan” also contain missing values. To treat the missing values, several steps were taken to make.

Firstly, missing states are filled using previous entries that have the same location. Next, missing station_code were filled using entries that have the same state and location as the entry missing a station_code. The location could not be filled because entries having the same station_code as the missing locations could not be found in the dataset.

Missing “temp” and “nitratenan” data is not filled to avoid having fabricated data affects the quality of data used for analysis.

Below is the dataset after the missing data from the “station_code” and “state” features have been filled.



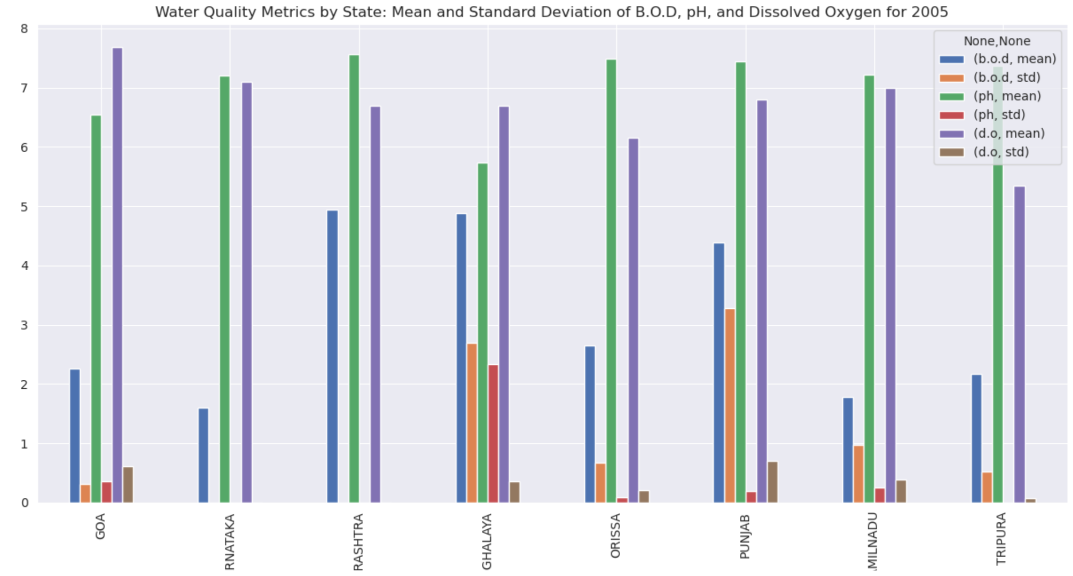
Handling duplicate values

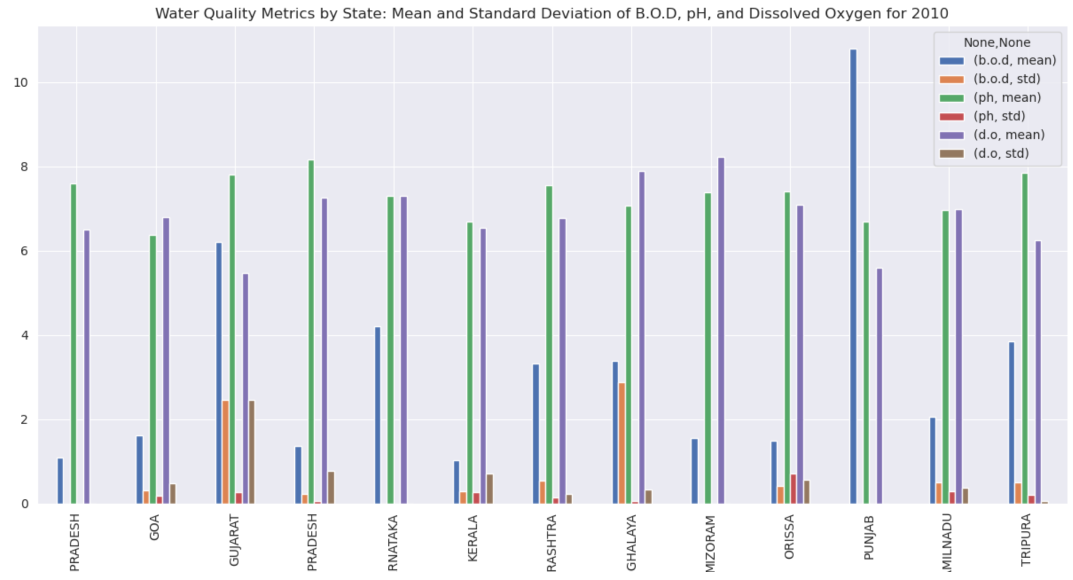
The dataset does not contain duplicate entries.

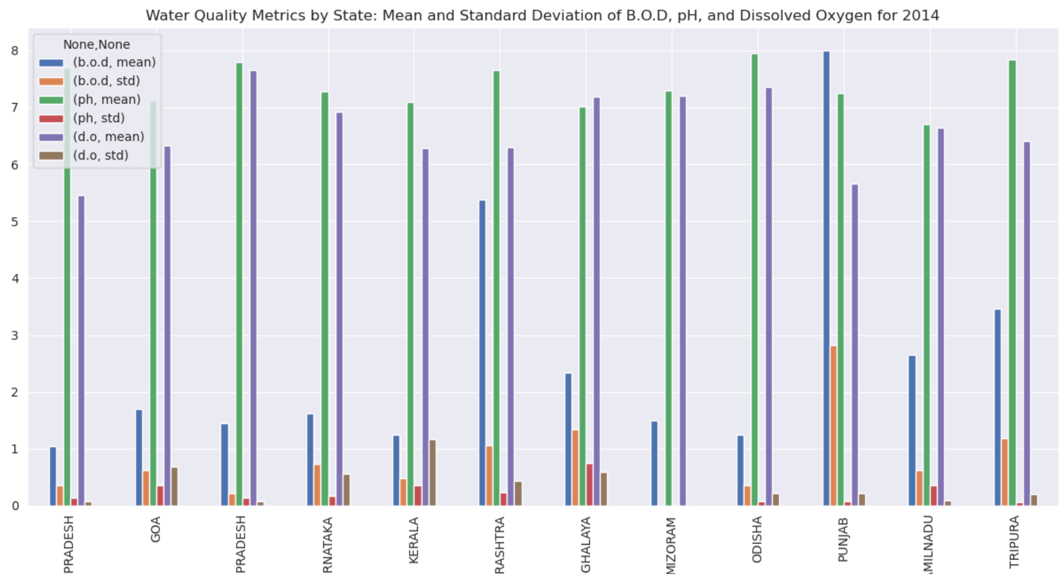
Exploratory Data Analysis (EDA)

Water Quality Metrics across the Years by State

Water quality metrics like Biological Oxygen Demand (B.O.D), Dissolved Oxygen (D.O) and PH affect water quality and can lead to pollution and reduce support for aquatic life. Safe ranges of pH for drinking water are from 6.5 to 8.5 for domestic use and living organisms' needs. Excessively high and low pH can be detrimental to the use of water. Dissolved oxygen (DO) is considered to be one of the most important parameters of water quality in streams, rivers, and lakes. The higher the concentration of dissolved oxygen, the better the water quality. The more organic material there is in the water, the higher the BOD used by the microbes will be. BOD is used as a measure of the power of sewage; strong sewage has a high BOD and weak sewage has low BOD.



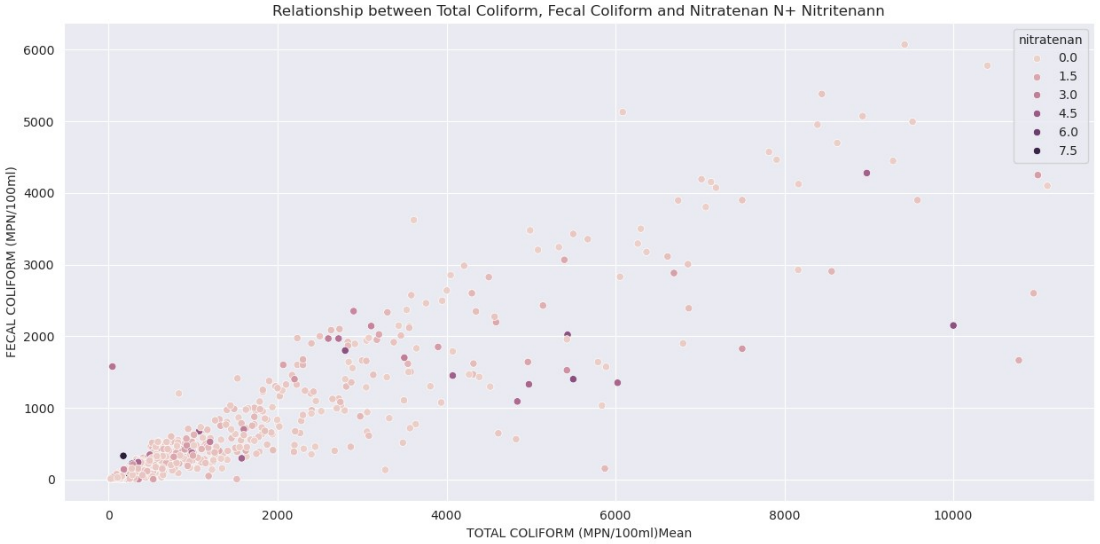




In 2004, fewer states were being monitored, with each state having relatively moderate pH, BOD and DO. BOD values range widely; generally, pristine waters have a value below 1 mgl−1, moderately polluted waters 2–8 mgl−1, and treated municipal sewage 20 mgl−1. With majority of the states have moderately polluted waters on average with high DO and a moderate pH. In 2014, most states showed BOD below 2 mgl-1 except Pashtra, Ghalaya, Punjab, Milnadu and Tripura. Punjab having the highest BOD is known for over-exploitation of groundwater for the agricultural process. These can account for the mineral and other organic materials that could be present in water bodies around Punjab.

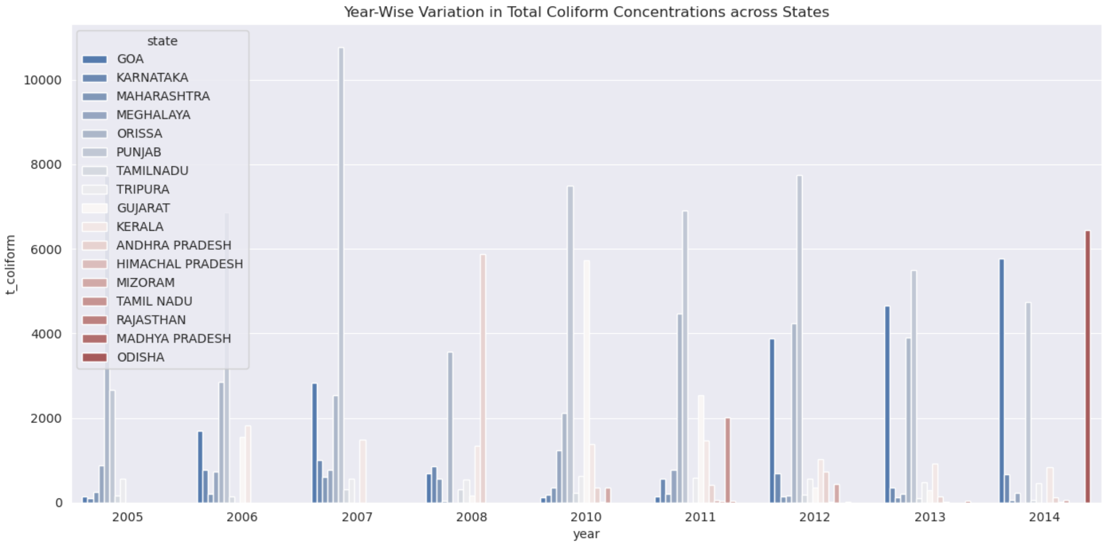
Relationship between Total Coliform, Fecal Coliform and NITRATENAN N+ NITRITENANN

Total Coliform and Fecal Coliform are bacterial groups used to assess water contamination by faecal matter. Total Coliform includes bacteria from various sources, while Fecal Coliform, a subset of Total Coliform, indicates recent pollution from faecal matter, posing an increased risk of waterborne diseases.



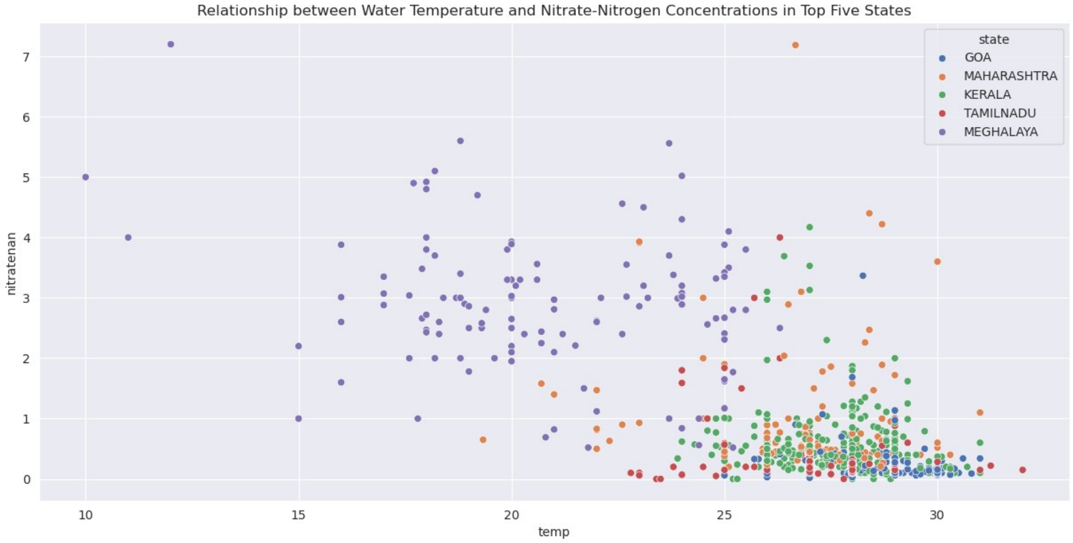
This shows that Fecal Coliform increases with an increase in Total Coliform (positive correlation). This makes sense since Fecal Coliform is a subset of Total Coliform. Fecal Coliform may reduce the nitrate, and it is also to be noted that higher nitrate indicates high Total Coliform bacteria as seen in the chart above with a few exceptions.

Total Coliform Mean By States from 2005 to 2014



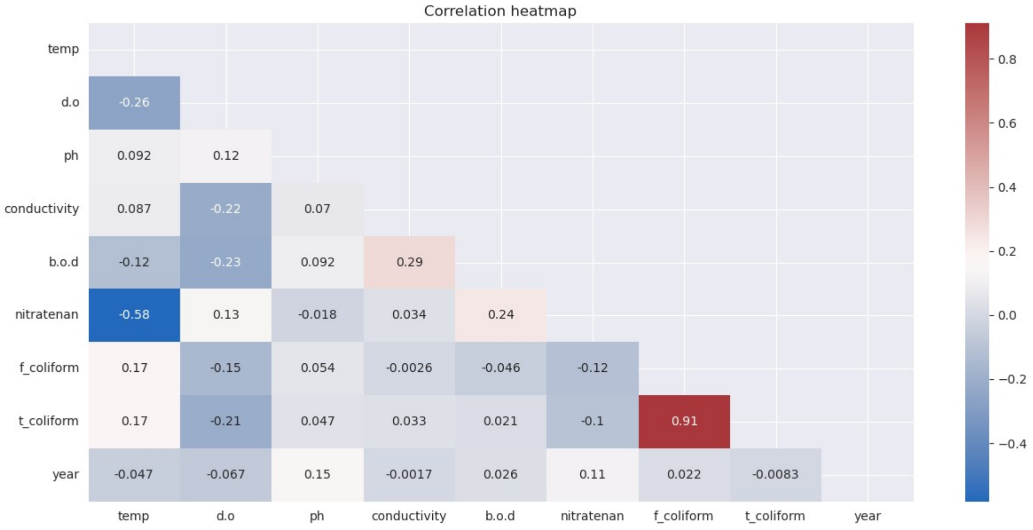
The chart shows a trend for the Total Coliform amount across states from 2005 to 2014. The Total Coliform concentration across states varies from year to year, with states like Goa and Punjab consistently having the highest numbers across several years. In 2014, Odisha had the highest Total Coliform. Factors like untreated sewage, industries and factors, and mining can be the cause of high Total Coliform concentration in these states.

Relationship between Water Temperature and Nitrate-Nitrogen Concentrations in Top Five States (By number of water stations)



The chart shows that entries for the Meghalaya state have lower temperatures and higher Nitrate-Nitrogen compared to the other top five states by number of stations. This shows a negative relationship between Temperature and Nitrate-Nitrogen.

Relationship between features of the dataset



	temp	d.o	ph	conductivity	b.o.d	nitratenan	f_coliform	t_coliform	year
temp									
d.o	-0.26								
ph	0.092	0.12							
conductivity	0.087	-0.22	0.07						
b.o.d	-0.12	-0.23	0.092	0.29					
nitratenan	-0.58	0.13	-0.018	0.034	0.24				
f_coliform	0.17	-0.15	0.054	-0.0026	-0.046	-0.12			
t_coliform	0.17	-0.21	0.047	0.033	0.021	-0.1	0.91		
year	-0.047	-0.067	0.15	-0.0017	0.026	0.11	0.022	-0.0083	

This shows that features of the dataset have weak correlation among each other, except Total Coliform and Fecal Coliform having a high positive correlation of 0.91, Temperature and Nitrate-Nitrogen having a moderate negative correlation of -0.58.

First step to model creation

The wrangle function is created to handle data preparation for model building. The function removes categorical variables like states, station_code and location. The function also removes extreme values using data between the 5th and 90th percentile of features containing extreme values. The function also fills in missing temperature values using the median temperature because the model data is skewed.