Quality of Water in Indian Water Bodies

Team Details:

Name: Aniket Aayush

SRN: PES1UG19CS061

Section: A

Name: Aakash G Acharya

SRN: PES1UG19CS005

Section: A

Name: Abhishek Aditya BS

SRN: PES1UG19CS019

Section: A

Name: A Sai Chaithanya

SRN: PES1UG19CS002

Section: A

Abstract

Through this project we wanted to analyse the Quality of different rivers and lakes in India, thereby predicting different levels of pollution in different water bodies in different states. For this we have selected a corresponding dataset and we have performed data analysis. Firstly, we have cleaned the dataset with appropriate methods and then we have performed certain visual and graph representation to infer important results, followed by normalisation of the cleaned data. Then we have used hypothesis testing on one of the population parameters to draw pre-conclusions on certain aspects, followed by heat map representing the correlation between two population parameters. In the end, we have used linear regression to show the dependence of two population parameters, followed by results and conclusion.

Introduction

Problem Statement at hand is to analyse the quality of water of different lakes and rivers in India to infer the pollution levels and using these results to improve the water quality. As we know, in India, water from rivers and lakes is used for daily activities, drinking, agriculture purposes majorly and for some other purposes. Since, with increasing population there is increase in the pollution in rivers and lakes, thereby making certain water bodies unfit for drinking and agriculture purposes as well.

This analysis will give some results with which we can predict and estimate about the water quality and hence, provide this data as important document to the government to improve the water quality by any measures possible and most importantly it can used to make our society and people around us aware of the certain dangers, and hence we can also make efforts reduce pollution of these water bodies. As of agriculture, farmers can know which water is better for them to use for their cultivation and be cautious and hence provide better and healthier yields.

Dataset

Combined data of historical water quality of certain locations in India. Pollutants measures in each column is the average values measured over a period of time. Source: "Indian government websites". The data set provides valuable information about important water qualifiers like TOTAL COLIFORM LEVEL, TEMPERATURE, pH, B.O.D, D.O etc. helpful in determining the water quality of various water bodies across India.
The Categorical columns are :

1. Station code : Geographical Location codes of various water bodies across India.
2. Locations : Physical Geographical locations of various water bodies across India.
3. State : various states across India.
4. Year ; The year in which the observations were made.

The Numerical columns are :

1. Temperature : Temperature levels of the water bodies
2. D.O. (mg/l) : Dissolved oxygen level in the water , lesser the D.O value more is the pollution level
3. pH : If pH < 7 the water is acidic , and is unfit for use
   If pH > 7 the water is alkaline and is safe
   Typically the water should be neutral with pH 7
4. Conductivity (µmhos/cm) : A sudden increase or decrease in conductivity in a body of water can indicate pollution.
5. B.O.D. (mg/l) : A low BOD is an indicator of good quality water, while a high BOD indicates polluted water.
6. Nitratenan N+ Nitratenann (mg/l) : Nitrates affect aquatic life. Nitrates have the same effect on aquatic plant growth as phosphates and thus the same negative effect on water quality. Thus higher the nitrate levels higher is the water contamination.
7. Fecal Coliform (mpn/100ml) : The presence of fecal coliform bacteria in aquatic environments indicates that the water has been contaminated with the fecal material of man or other animals.Water pollution caused by fecal contamination is a serious problem due to the potential for contracting diseases from pathogens (disease-causing organisms).

8. Total Coliform (mpn/100ml)Mean : Total coliforms are a variety of bacteria, parasites, and viruses, known as pathogens, can potentially cause health problems if humans ingest them. EPA considers total coliforms a useful indicator of pathogens in water.So higher the total coliform levels more is the pathogens in water hence more is the water contamination.

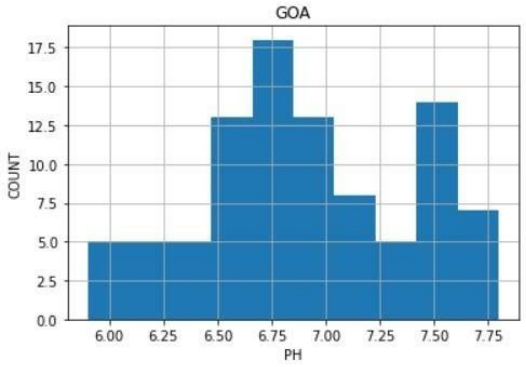
Preprocessing or Data Cleaning

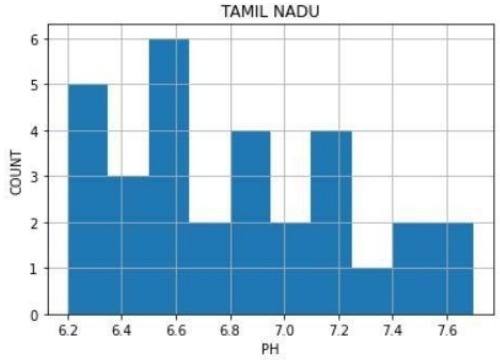
We have cleaned the data in our dataset by first **checking for any duplicate rows and removing** them as they wouldn't help our analysis and then **replacing all the empty values(NaN)** of each parameter or water qualifiers by the **mean of data of that parameter(column)** corresponding to water body locations in that particular state.Then we **fixed the incorrect capitalization** and **typos in the state names** and converted the numerical columns to numeric type and **converted the year to pandas date time format**,and **station code to String type** since it belongs to categorical data type.We have also removed certain number of rows corresponding to a particular state as most of the values are NaN or missing, hence are not useful or cannot provide valuable information. By using boxplot visualisation, we have identified certain outliers in our data. So, to **remove these outliers** we have used the concept of minimum value and maximum value of a boxplot calculated by certain formulas and any value less or greater compared to the former is removed. Now, we have a clean data set with no missing values and with all outliers removed making our further analysis more precise and accurate.

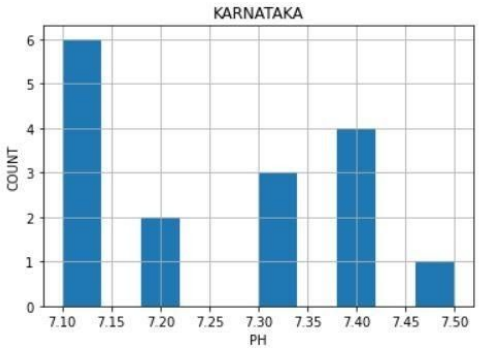
Exploratory Data Analysis

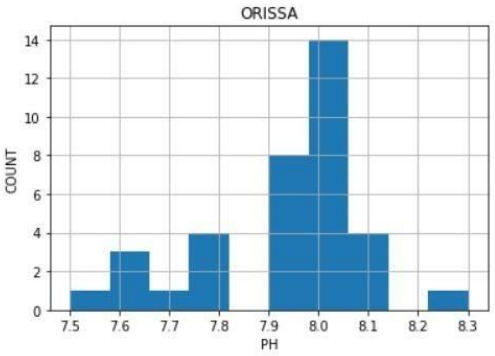
We have plotted graphs corresponding to avg. pH value in each state. We have used **Histograms** as the visualisation tool which indicates avg. pH for all states, infer conclusions. pH value less than 7 would indicate the water is acidic, hence making not safe for useful for domestic or agricultural activities. We have inferred the following,

- **Tamil Nadu** has most water bodies with **pH level 6.5**
- Followed by **Goa** which has most water bodies with **pH level 6.75**
- Typically, **India water is neutral between 7 and 8**, so all other states other than Tamil Nadu, Goa and Orissa have pH levels between 7 and 8, water is neither acidic nor basic and perfect for all uses.
- **Ph value greater 8** means water is **alkaline** and is not harmful and useful for domestic as well as industrial purposes.
- So, **Orissa** has the most water bodies with **pH level 8.05.**



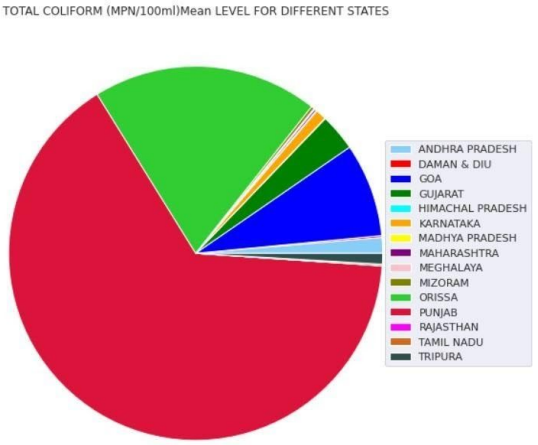


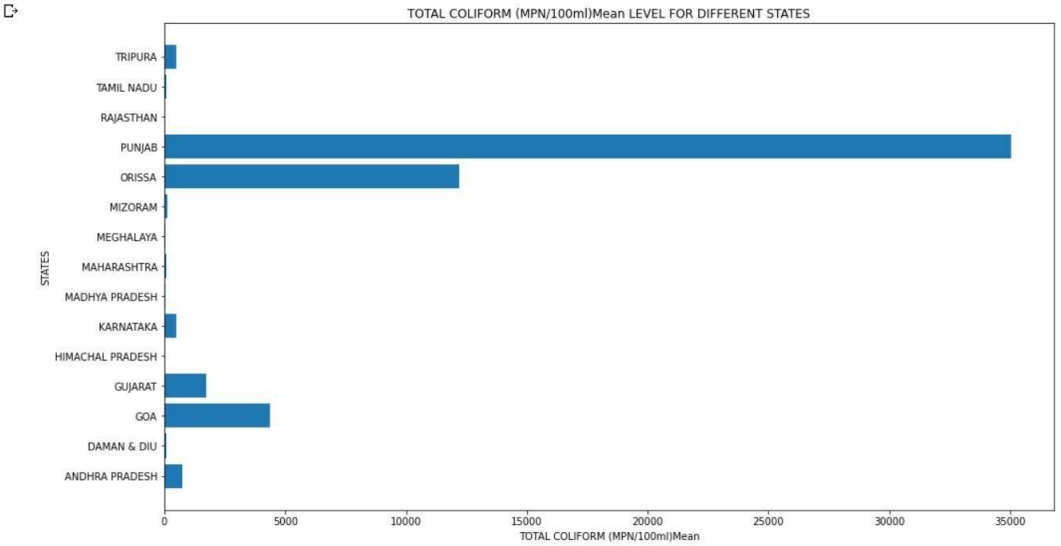




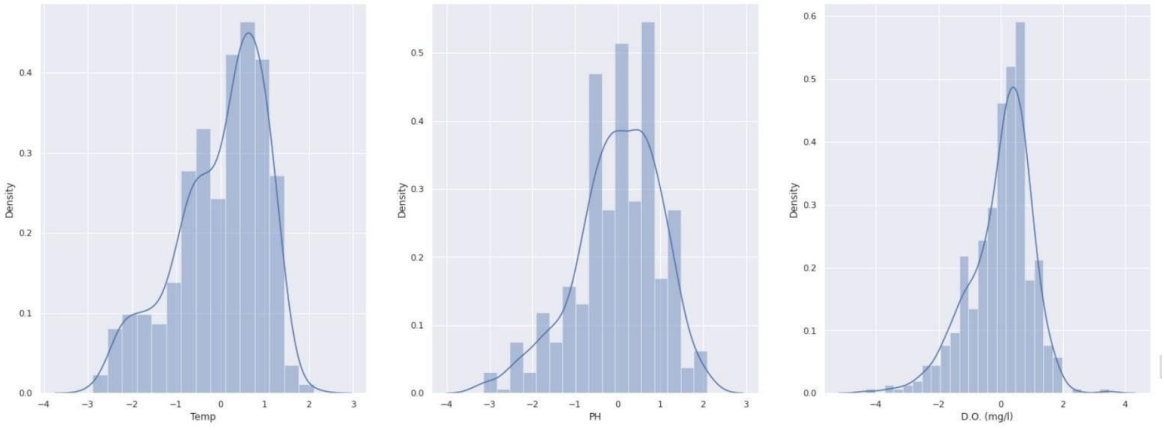
We have also used **Bar Charts and Pie Charts** to represent sensible and important statistics and draw valuable results. Following is the inference,

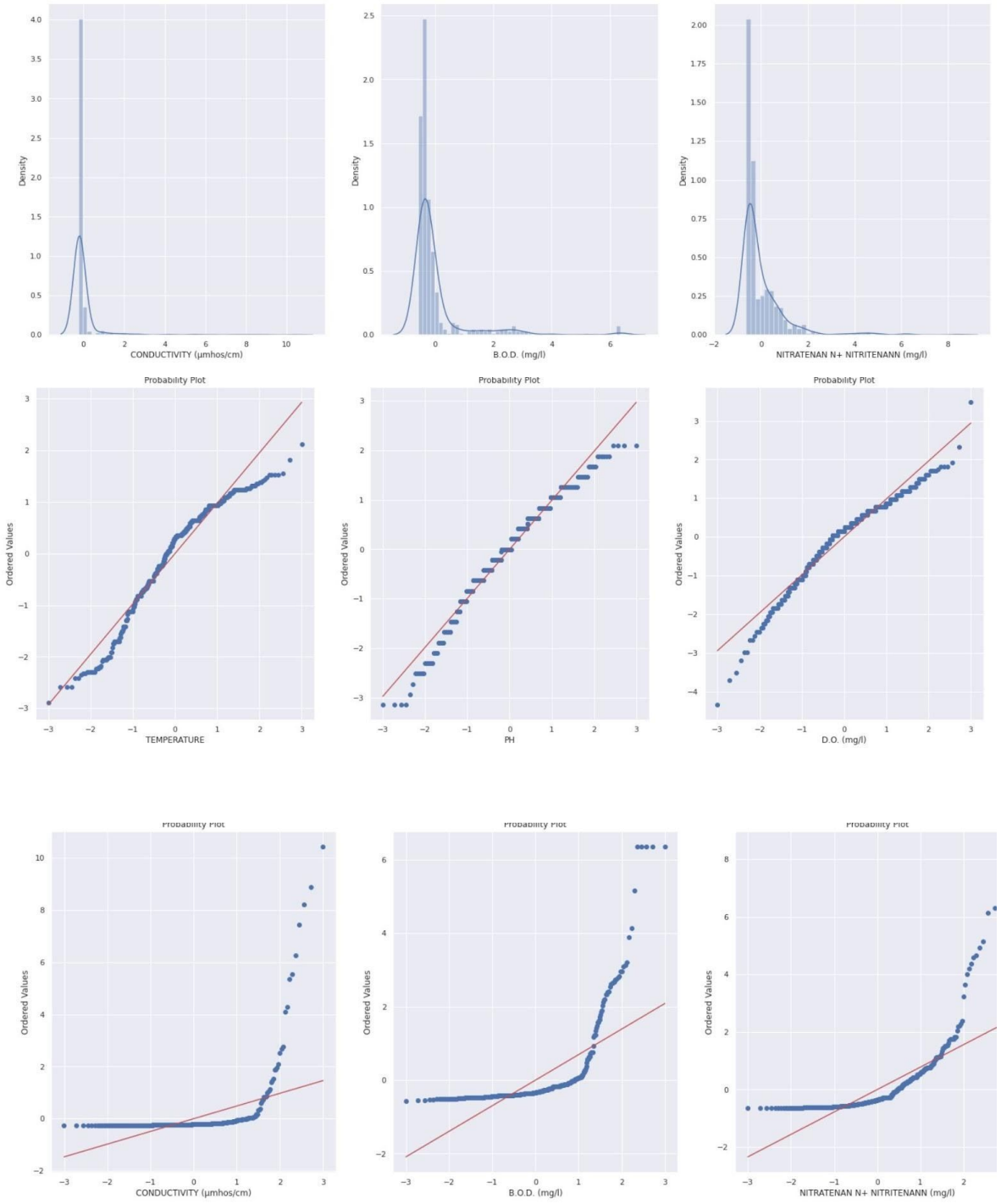
- **PUNJAB** has the Highest Total Coliform Level and Second Highest B.O.D levels thereby showing that it has the **MOST POLLUTED WATER IN INDIA.**
- **RAJASTHAN and HIMACHAL PRADESH** has lowest Total Coliform levels thereby they have the **LEAST POLLUTED WATER IN INDIA.**





Now, after this we have normalised our numerical column data to reduce the mean to zero and standard deviation to one ,to bring the values to a common scale . Using this normalised data, we have plotted Distribution plots and Normal Probability plots. As you can see **we have both Normal Distribution and Right skewed Distribution.**





Hypothesis Testing

B.O.D(Biological Oxygen Demand) column is used for hypothesis testing. Here, mean of the B.O.D of the population is calculated. We check if the same applies to the sample. Null Hypothesis is population mean equals 4.5mg/l and the alternate hypothesis is population mean not equals 4.5 mg/l (values corresponding to a given random sample).

We generate a random sample from the population and perform the Z Statistic test on this sample corresponding to the B.O.D column. Using the z-score value we determine the p – value and then compare the obtained p value to alpha. In our analysis, for the given random sample we
p value is 0.26 and alpha is 0.05. Since, p-value is greater than alpha we conclude that,
We Fail to Reject NULL Hypothesis.

Results and Discussion

1. **PUNJAB** has the highest Total coliform level and second highest B.O.D levels thereby showing that it has the most polluted water in India, **Rajasthan and Himachal Pradesh** having the least polluted water bodies.
2. **Daman & Diu** has the highest B.O.D (mg/l) level so we can say that Daman & Diu water bodies has very less oxygen saturation in water due to the presence of algae pathogens, toxic substances, proving harmful for the aquatic life.
3. **TAMIL NADU** has most water bodies with pH level 6.5 followed by **GOA** which has most water bodies with pH level 6.75 means water is **acidic** and is unfit for use and harmful for the environment as well.
4. We fail to reject the null Hypothesis which states B.O.D mean is 4.5 mg/l
5. **B.O.D and Total coliform** & **Fecal coliform and Total coliform levels** are positively correlated qualifiers.