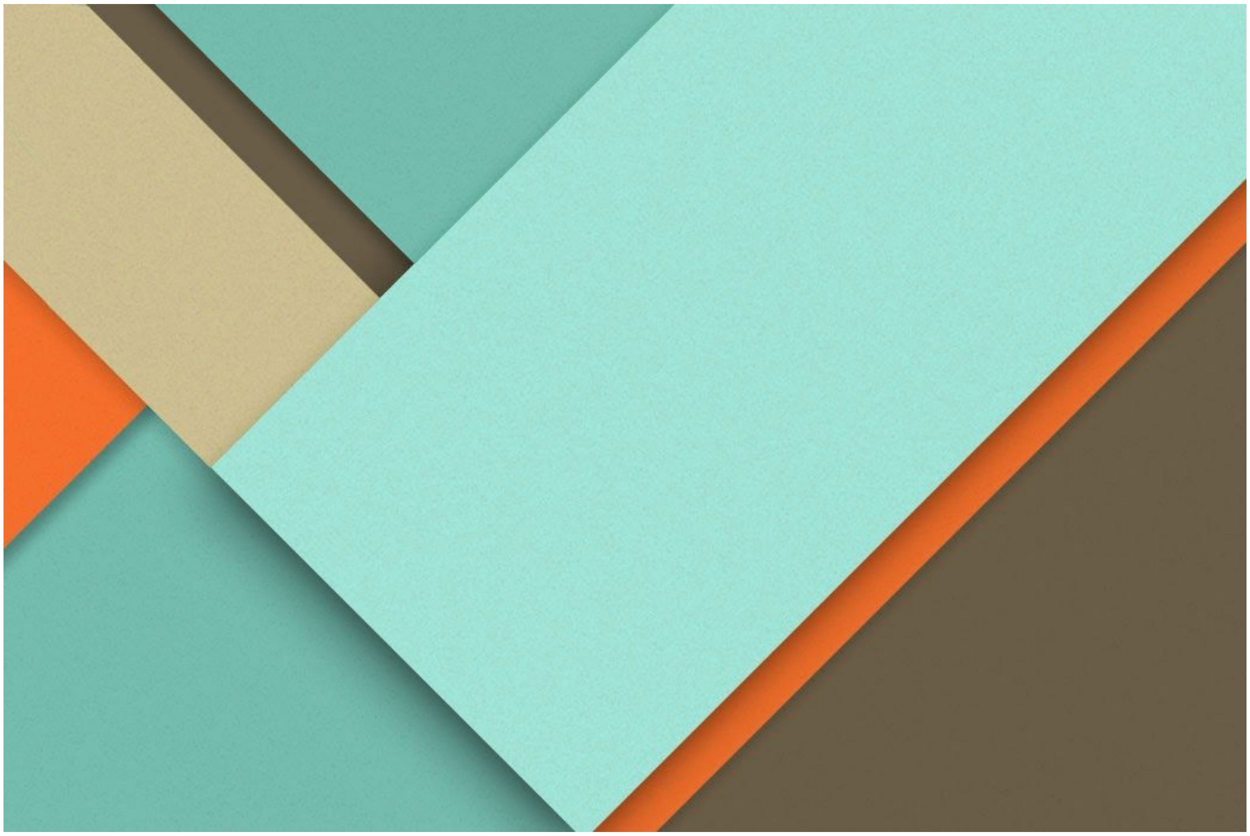


BTP-2

12.12.2020

WQI Calculation using Water Quality Dataset

- Divyesh Jain
- Sidharth Jain
- Akshay Kharbanda
- Sarthak Singh
- Parth Garg

Abstract

In this paper, we are analyzing the water Quality data collected by CPCB. We did data preprocessing in this work as it is the most crucial task .We did normalization, binning and discretization. Our data set contains water quality data from monitoring stations, located on all important rivers, lakes including wells for groundwater assessment form the basis of the National Water Monitoring Program (NWMP).

Introduction

In this work we did data pre-processing . Data cleaning is an important task in data mining because the data has to be clean, noise free and removing missing values. Every machine learning algorithm will be operated on the dataset. So we will get better insight when data cleaning techniques are applied. Results achieved from the machine learning algorithms will not be correct if operated on data without enough cleaning. In this work we discretize continuous data using unsupervised classification techniques such as binning, etc.

Data Preprocessing

For the preprocessing we have used 3 techniques, which are Data Cleaning, Data Reduction and Data Transformation.

A. Data Cleaning

a. **Missing Values** - We replaced all the missing values with the column means and also removed the records which have erroneous data values.

b. **Noisy Data** - we have replaced missing values by the column mean and the rows with more than 40% missing data were removed from the dataset.we have also removed discrepancies in the names of states which were initially more than 28.

B. Data Reduction

a. **Dimensionality Reduction** - We have used PCA for dimensionality reduction, which includes calculating eigenvalues and eigenvectors in reducing features and getting data points in new dimensions which would be helpful in displaying any underlying qualities of the data.

b. **Numerosity Reduction** - We have used Sampling for numerosity reduction, individual datasets with higher amounts of data have been sampled accordingly in a random manner for getting only the required amount of data.

C. Data Transformation

a. **Normalization** - We have performed Normalization to allow effective processing of the data. We have used Min-Max Normalization with 0 as min and 1 as max in order to scale down the values between 0-1 so that no specific feature has more effect due to larger values in further processing.

b. **Aggregation** - We have used Aggregation to replace numerous features into a new single feature which captures all the patterns of old features. we have replaced min and max columns of temperature by a single new column which represents the range.

Water Quality Index (WQI)

WQI

It may be defined as a rating, reflecting the composite influence of different water quality parameters on the overall quality of water. The main objective of computing of water quality index (WQI) is to turn the complex water quality data into information which is easily understandable and usable.

Selection of parameters

First, study the Indian Standard (BIS 2004) for drinking water specification. Here, the physicochemical parameters along with the desirable limits and related health effects are given. A parameter has to be selected based on its impact in the overall quality of water and health effects.

Computation of WQI

The WQI is computed following the three steps.

First step – Assigning of weight (wi) to the selected water parameters (e.g., pH, TDS, TH, HCO3, Cl, SO4, NO3, Fe, ……) according to their relative importance in the overall quality of water for drinking purposes (weight may be from 1 to 5).

Second step – Computation of a relative weight (Wi) of the chemical parameter using the following equation:

$$Wi = wi / \sum wi \ (i = 1 \text{ to } n)$$

where, Wi is the relative weight, wi is the weight of each parameter and 'n' is the number of parameters

Third step - Assigning of a quality rating scale (qi) for each parameter, as below:

$$qi = (Ci / Si) \times 100$$

where, qi is the quality rating, Ci is the concentration of each chemical parameter in each water sample in mg/l, and Si is the guideline value/desirable limit as given in Indian drinking water standard (BIS 2004).

For computation of WQI, the sub index (SI) is first determined for each chemical parameter, as given below:

$$SIi = Wi \times qi$$

$$WQI = \sum SIi \ 1 - n$$

where, SIi is the subindex of ith parameter; Wi is relative weight of ith parameter; qi is the rating based on concentration of ith parameter and 'n' is the number of chemical parameters.

Classification of water

The water may be classified into five types based on computed WQI as given below:

WQI range	**Water Type**
< 50	Excellent water
50 - 100	Good water
100- 200	Poor water
200 – 300	Very poor water
> 300	Water unsuitable for drinking

Data Visualization

We have obtained the following insights, in the form of bar graphs after the pre-processing of the data.

1. State vs pH
2. State vs BOD
3. State vs Coliform
4. State vs DO

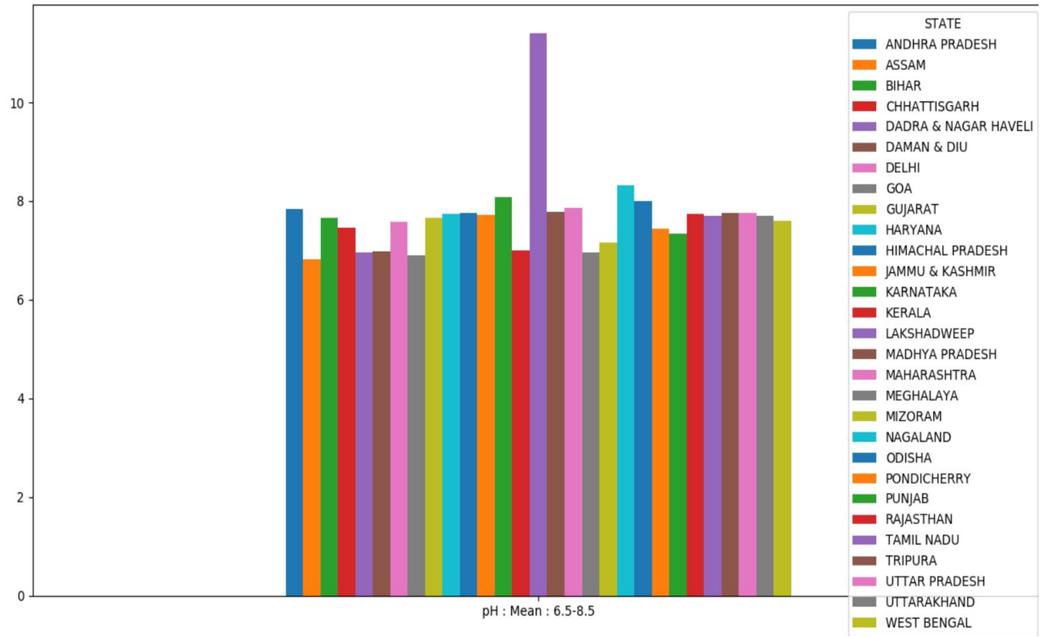


Fig1. State vs pH

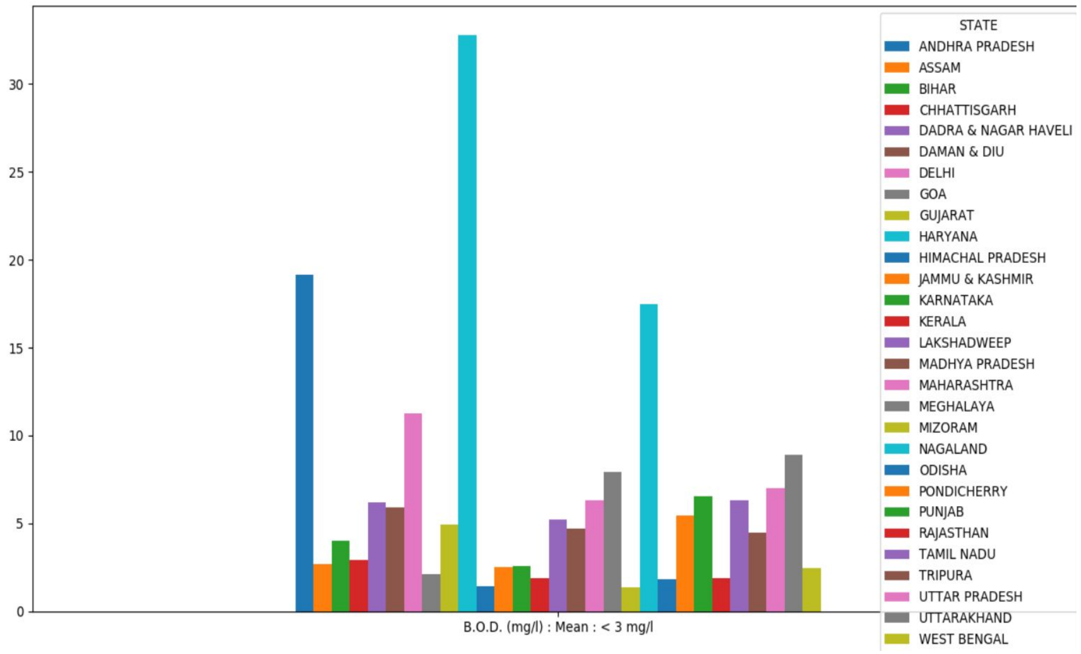


Fig2. State vs B.O.D

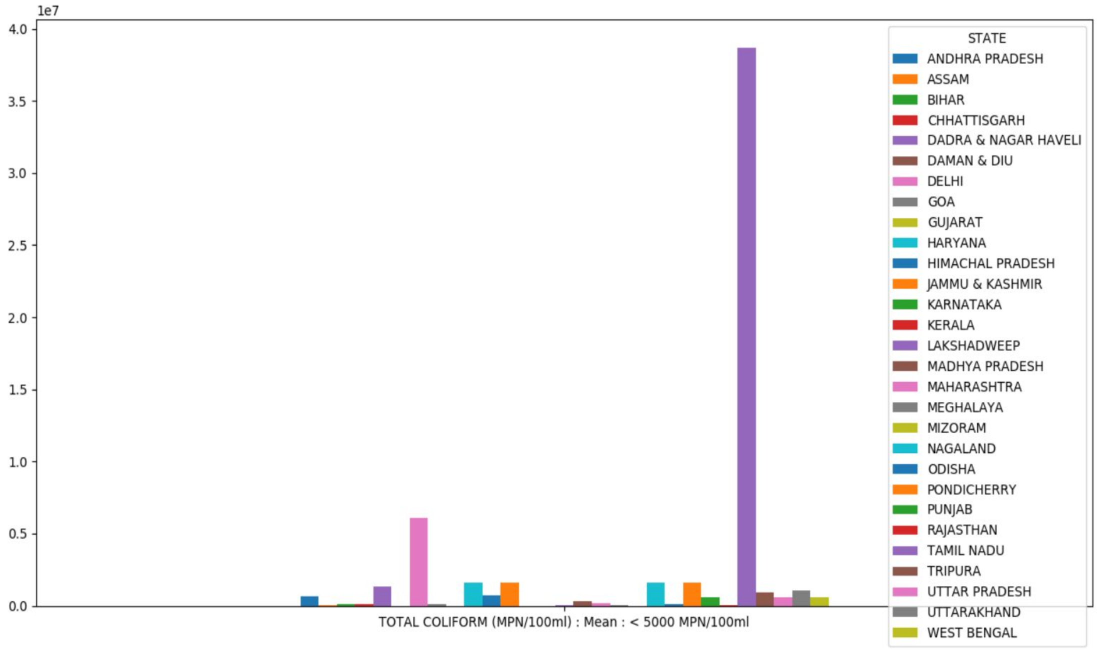


Fig3. State vs Coliform

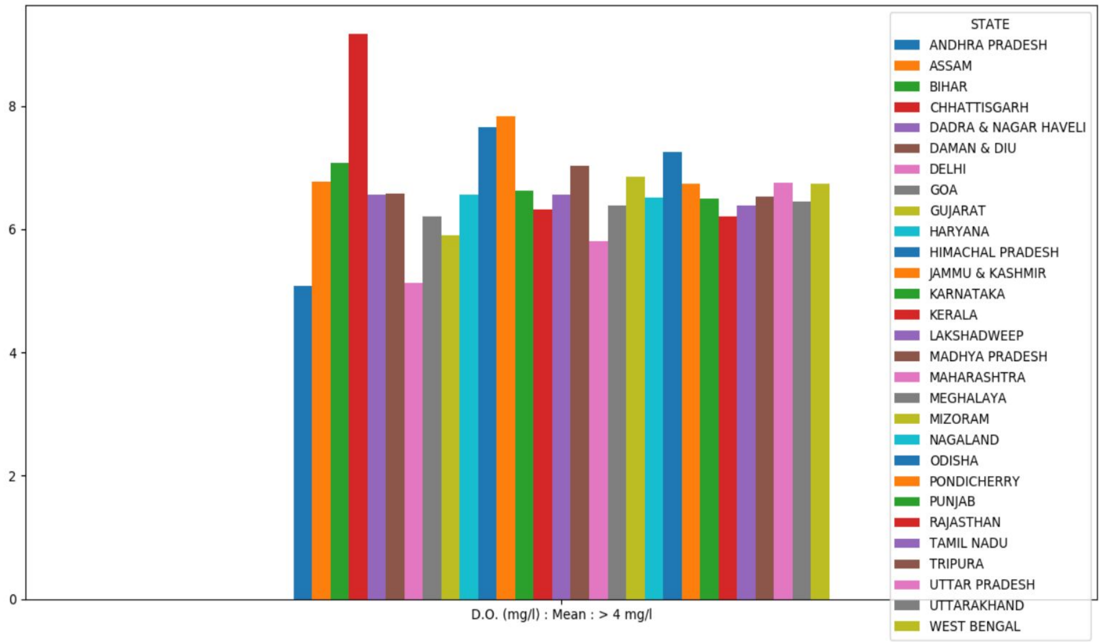
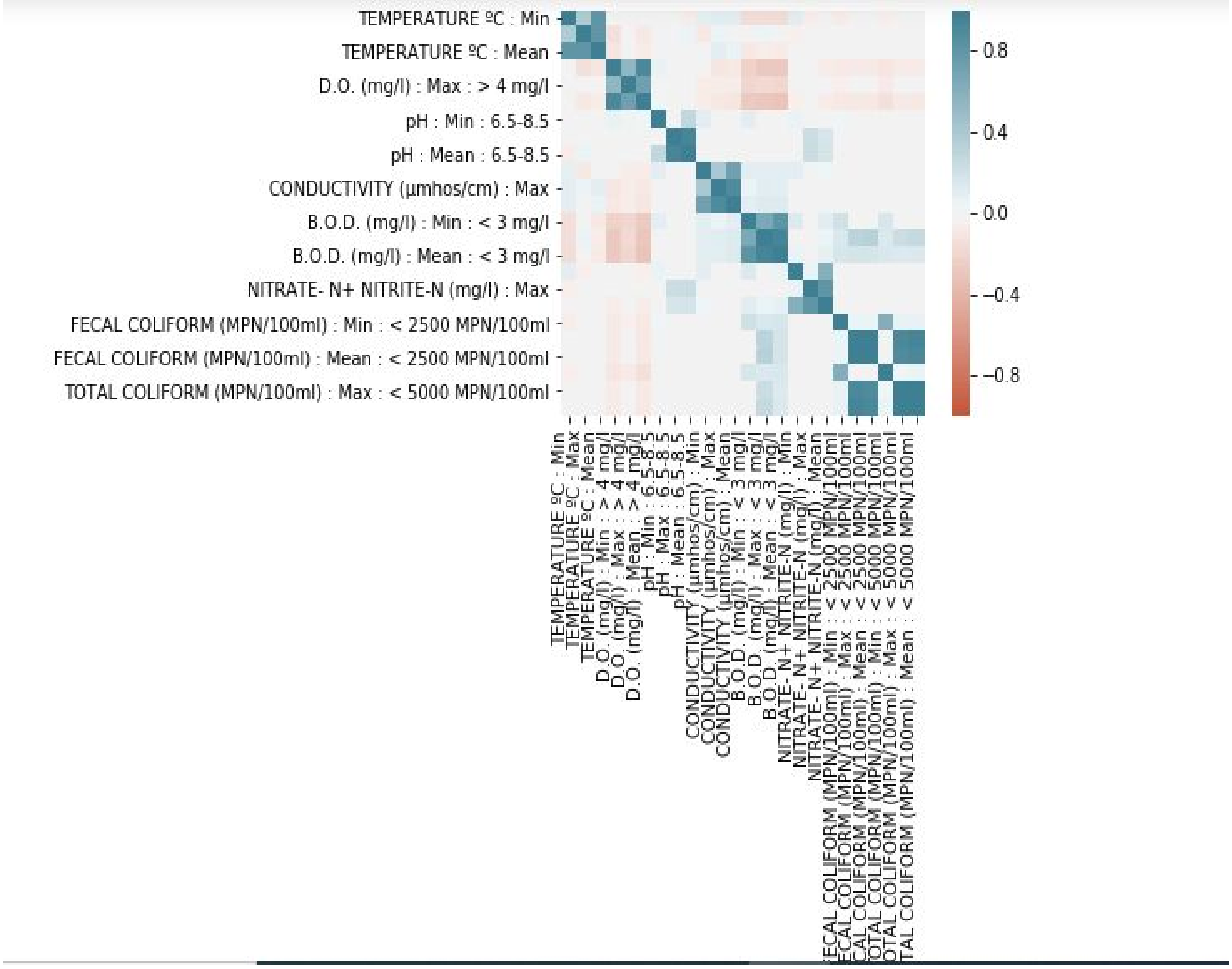


Fig4. State vs DO

We also obtained a correlation plot between features:


TEMPERATURE ºC : Min
TEMPERATURE ºC : Mean
D.O. (mg/l) : Max : > 4 mg/l
pH : Min : 6.5-8.5
pH : Mean : 6.5-8.5
CONDUCTIVITY (µmhos/cm) : Max
B.O.D. (mg/l) : Min : < 3 mg/l
B.O.D. (mg/l) : Mean : < 3 mg/l
NITRATE- N+ NITRITE-N (mg/l) : Max
FECAL COLIFORM (MPN/100ml) : Min : < 2500 MPN/100ml
FECAL COLIFORM (MPN/100ml) : Mean : < 2500 MPN/100ml
TOTAL COLIFORM (MPN/100ml) : Max : < 5000 MPN/100ml
TEMPERATURE ºC : Min
TEMPERATURE ºC : Max
TEMPERATURE ºC : Mean
D.O. (mg/l) : Min : > 4 mg/l
D.O. (mg/l) : Max : > 4 mg/l
D.O. (mg/l) : Mean : > 4 mg/l
pH : Min : 6.5-8.5
pH : Max : 6.5-8.5
pH : Mean : 6.5-8.5
CONDUCTIVITY (µmhos/cm) : Min
CONDUCTIVITY (µmhos/cm) : Max
CONDUCTIVITY (µmhos/cm) : Mean
B.O.D. (mg/l) : Min : < 3 mg/l
B.O.D. (mg/l) : Max : < 3 mg/l
B.O.D. (mg/l) : Mean : < 3 mg/l
NITRATE- N+ NITRITE-N (mg/l) : Min
NITRATE- N+ NITRITE-N (mg/l) : Max
NITRATE- N+ NITRITE-N (mg/l) : Mean
ECAL COLIFORM (MPN/100ml) : Min : < 2500 MPN/100ml
ECAL COLIFORM (MPN/100ml) : Max : < 2500 MPN/100ml
CAL COLIFORM (MPN/100ml) : Mean : < 2500 MPN/100ml
OTAL COLIFORM (MPN/100ml) : Min : < 5000 MPN/100ml
OTAL COLIFORM (MPN/100ml) : Max : < 5000 MPN/100ml
TAL COLIFORM (MPN/100ml) : Mean : < 5000 MPN/100ml