The Chemical Analysis of Water Quality of India

Sonali Alpeshbhai Thakkar,
Assistant Professor, Silver Oak College of Computer Application,
Silver Oak University, Ahmadabad, Gujarat

Abstract:- Water quality is affected to animal environment as well as human life so the current data set is analyzing the factor which is affecting to water quality. The important parameters of water quality are ph, temperature, conductivity, dissolved oxygen, biodisolved oxygen, Biochemical oxygen demand, nitrate and chloroform. Water quality refers to the chemical, physical, biological and radiological characteristics of India. It is a measure of the condition of water relative to the requirements of one or more biotic species and or to any human need or purpose. It is most frequently used by reference to a set of standards against which compliance, generally achieved through treatment of the water, can be assessed.

Keywords:- *Water Quality, Chemical Analysis, Data Visualization, Oxygen.*

I. INTRODUCTION

Water is an essential nutrient and plays a key role in the human body. We can survive up to several weeks without food, but only a few days without water. Every system in the body, from cells and tissues, to vital organs requires water to function. All plants and animals need water to survive. There can be no life on earth without water. Why is water so important? Because 60 percent of our body weight is made up of water. Our bodies use water in all the cells, organs, and tissues, to help regulate body temperature and maintain other bodily functions. Because our bodies lose water through breathing, sweating, and digestion.

II. PROBLEM STATEMENT

- Which areas in India have a lot of water quality degradation issues over the years?
- Which chemical is predominantly present in most of the water quality issues?
- If there are any associations between the water quality data and the other developmental data. If there is, then what is the extent (visualization) and how can we address it?
- If there are any repetitive patterns of water quality degradation in the same area for multiple years.
- As a whole, for the country, is the water quality degrading or upgrading (number of instances reported of water quality getting affected)?

➢ *Objectives of the Chemical Water Analysis*

➢ The objective here is to

- To identify a which factor is affected to water quality
- To identify a which chemical is important to water quality
- Find out PH variation on each state so we can easily find out what are the sea-animals survival on each state
- To find out where dissolved oxygen is going down
- Find out Mineral Quality of water
- To find out what are the required minerals for drinking water
- To identify a which factor is affected to water quality

III. DATA ORGANIZATION

➢ *Data Volume*

The Dataset contains 1992 rows and 12 columns.

➢ *Data Summary*

```
c<-read.csv("D:\\SEM 8\\PROJECT\\project_graphs\\water_dataX_clean.csv") summary(c)
```

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAY2480

Table 1: Sample Data Table

Attributes	**No of null Values**	**Min**	**Max**	**1IQR**	**3IQR**	**Median**	**Mean**	**Outliners**
STATION CODE	YES	NA						NA
LOCATIONS	YES	NA	NA	NA	NA	NA	NA	NA
STATE_OLD	YES	NA	NA	NA	NA	NA	NA	NA
STATE	YES	NA	NA	NA	NA	NA	NA	Yes
Temp	NA	0.000	35.0	24.0	28.3	27.0	25.0	NA
D.O. (mg/l)	NA	0.000	11.400	5.900	7.200	6.700	6.293	NA
PH	NA	0.000	67115.0	6.9	7.7	7.3	111.6	Yes
CONDUCTIVITY (瘦 hos/cm)	NA	0.000	65700.0	76.0	568.5	180.0	1764.0	Yes
B.O.D. (mg/l)	NA	0.000	534.50	1.10	3.60	1.80	6.79	Yes
NITRATE N+ NITRITEN (mg/l)	NA	0.000	108.70	0.15	1.20	0.43	1.44	Yes
FECAL COLIFORM (MPN/100ml)	YES	0.000	272521616	7	628	124	304991	NA
TOTAL COLIFORM (MPN/100ml)Mean	NA	0.000	511090873	73	1696	394	498305	Yes
Year	NA	2003	2014	2008	2013	2011	2010	Yes

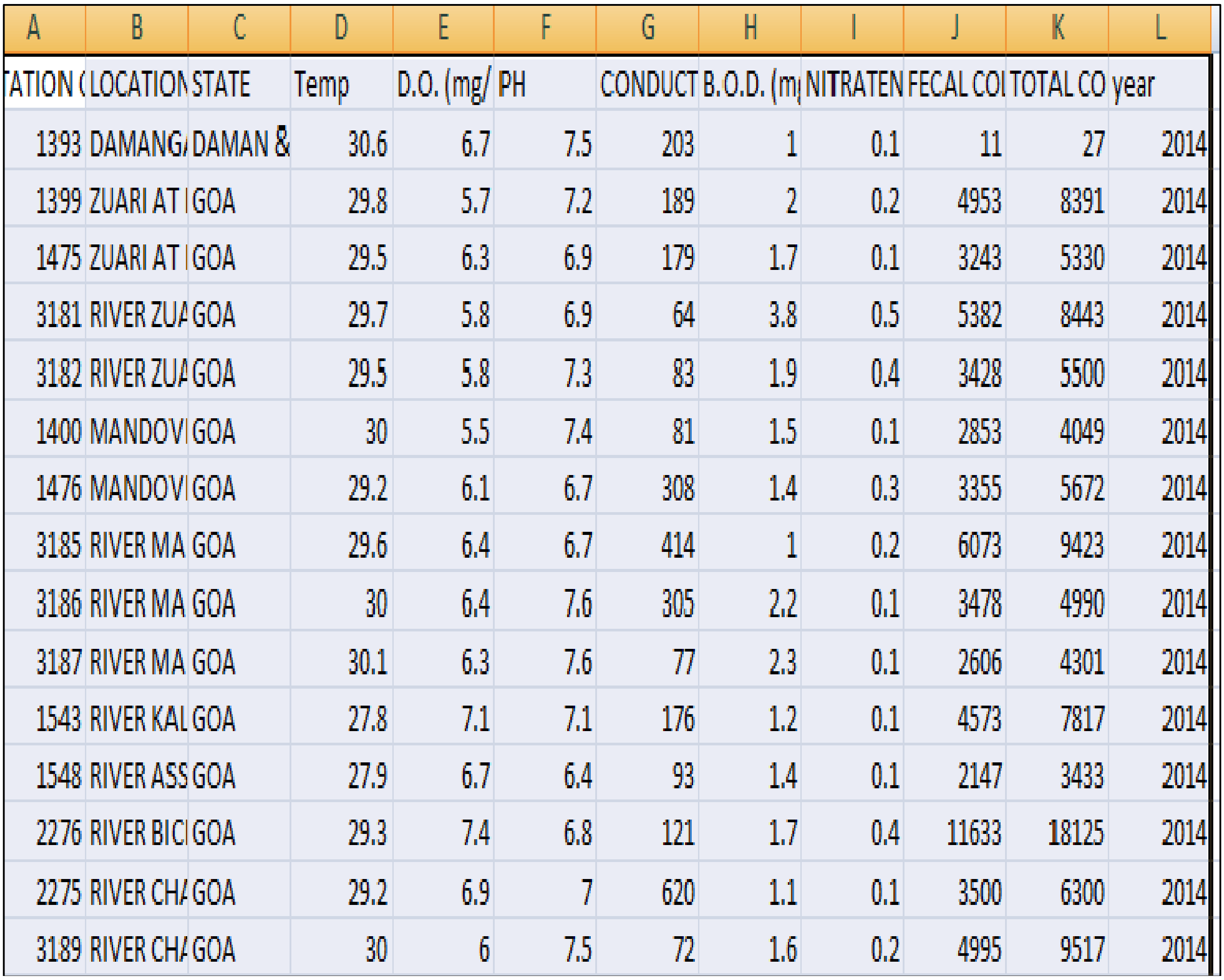
Table 2: Data Visualization

Box Plot	Is the visual representation of the depicting groups of numerical data .A box plot consisting of 5 things. 1) Minimum 2) First Quartile or 25% 3) Median Second Quartile) or 50% 4) Third Quartile or 75% 5) Maximum.
Histogram	A histogram is a plot of the frequency distribution of numeric array by splitting it into small equal sized bins. We can investigate the distributions of the data by reviewing histograms. 1) A histogram for each variable in the time series. 2) A histogram of active power consumption for the two full years of data.
Scatter plot	Scatterplots show many points plotted in the Cartesian plane. Each point represents the values of two variables. One variable is chosen in the horizontal axis and another in the vertical axis. A scatterplot is a useful way to visualize the relationship between two variables. Similar to correlations, scatterplots are often used to make initial diagnoses before any statistical analyses are conducted.

IV. DATA CLEANING

For creating a predictive model firstly the data must be clean and accurate and null values should not be there.

Table 3: Raw Data



A	B	C	D	E	F	G	H	I	J	K	L
TATION (	LOCATION	STATE	Temp	D.O. (mg/	PH	CONDUCT	B.O.D. (m	NITRATEN	FECAL COI	TOTAL CO	year
1393	DAMANG/	DAMAN &	30.6	6.7	7.5	203	1	0.1	11	27	2014
1399	ZUARI AT I	GOA	29.8	5.7	7.2	189	2	0.2	4953	8391	2014
1475	ZUARI AT I	GOA	29.5	6.3	6.9	179	1.7	0.1	3243	5330	2014
3181	RIVER ZUA	GOA	29.7	5.8	6.9	64	3.8	0.5	5382	8443	2014
3182	RIVER ZUA	GOA	29.5	5.8	7.3	83	1.9	0.4	3428	5500	2014
1400	MANDOVI	GOA	30	5.5	7.4	81	1.5	0.1	2853	4049	2014
1476	MANDOVI	GOA	29.2	6.1	6.7	308	1.4	0.3	3355	5672	2014
3185	RIVER MA	GOA	29.6	6.4	6.7	414	1	0.2	6073	9423	2014
3186	RIVER MA	GOA	30	6.4	7.6	305	2.2	0.1	3478	4990	2014
3187	RIVER MA	GOA	30.1	6.3	7.6	77	2.3	0.1	2606	4301	2014
1543	RIVER KAL	GOA	27.8	7.1	7.1	176	1.2	0.1	4573	7817	2014
1548	RIVER ASS	GOA	27.9	6.7	6.4	93	1.4	0.1	2147	3433	2014
2276	RIVER BICI	GOA	29.3	7.4	6.8	121	1.7	0.4	11633	18125	2014
2275	RIVER CHA	GOA	29.2	6.9	7	620	1.1	0.1	3500	6300	2014
3189	RIVER CHA	GOA	30	6	7.5	72	1.6	0.2	4995	9517	2014

➢ *In Used Water Quality Data Columns Like:*

- PH
- D.O. (mg/l)
- CONDUCTIVITY
- Temp
- B.O.D.
- FECAL COLIFORM
- NITRATENAN

Have missing values. Depending on the model, these columns can be removed completely. However, these columns could be important for the model. So, the records that have null values can be removed.

➢ *Techniques used for Cleaning Data*

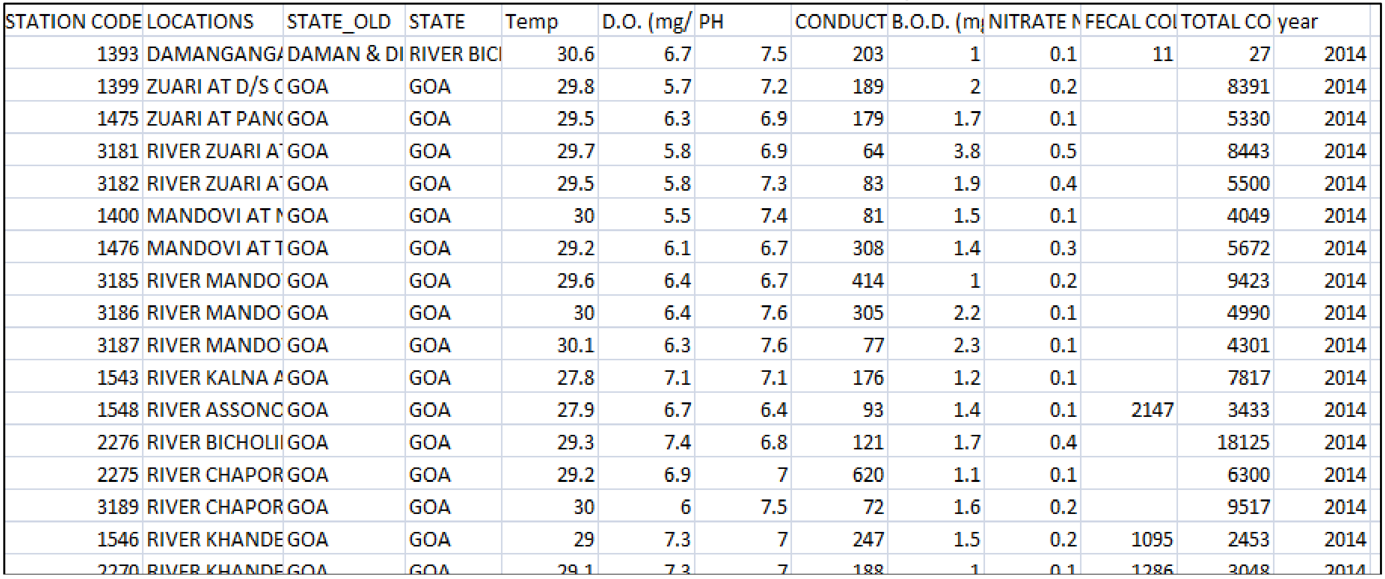
- Excel's in build filtering method is used for cleaning the data.
- Python is used for cleaning the null value and filling the missing value.
- Missing data is filled by the mean method.
- Wrong data is removed by filtering using Excel.
- After cleaning it containing 639624 examples, 12 attributes and some missing values and null values.

- *Python Code for Cleaning*

```
import csv
fp=open('water_dataX.csv','r')
fp1=open('water_dataX_cleaned.csv','w')
mywriter = csv.writer(fp1)
row1=[]
temp=0
reader = csv.reader(fp)
for row in reader:
    #print(row)
    #row=line.split(',')
    if temp==0:
        print(row)
        mywriter.writerow(row)
        temp=1
        #next(reader, None)
    else:
        #print(row)
        row1.append(row[0])
        row1.append(row[1])
        row1.append(row[2])
        if float(row[3])<30:
            row1.append(0)
        else:
            row1.append(row[3])
        if float(row[4])<7.6 or float(row[4])>14.6:
            row1.append(0)
        else:
            row1.append(row[4])
        if float(row[5])<4 or float(row[5])>8.5:
            row1.append(0)
        else:
            row1.append(row[5])
        if float(row[6])<50 or float(row[6])>1500:
            row1.append(0)
        else:
            row1.append(row[6])
        if float(row[7])<2 or float(row[7])>8:
            row1.append(row[7])
        else:
            row1.append(0)
        if float(row[8])>5:
            row1.append(0)
        else:
            row1.append(row[8])
        if float(row[9])>2500:
            row1.append(0)
        else:
            row1.append(row[9])
        row1.append(row[10])
        row1.append(row[11])
        mywriter.writerow(row1)
        row1=[]
        #next(reader, None)
fp.close()
fp1.close()
```

- *After Cleaning*

Table 4: Data Information After Cleaning



STATION CODE	LOCATIONS	STATE_OLD	STATE	Temp	D.O. (mg/	PH	CONDUCT	B.O.D. (mɩ	NITRATE N	FECAL COI	TOTAL CO	year
1393	DAMANGANG/	DAMAN & DI	RIVER BICI	30.6	6.7	7.5	203	1	0.1	11	27	2014
1399	ZUARI AT D/S (	GOA	GOA	29.8	5.7	7.2	189	2	0.2		8391	2014
1475	ZUARI AT PAN(	GOA	GOA	29.5	6.3	6.9	179	1.7	0.1		5330	2014
3181	RIVER ZUARI A'	GOA	GOA	29.7	5.8	6.9	64	3.8	0.5		8443	2014
3182	RIVER ZUARI A'	GOA	GOA	29.5	5.8	7.3	83	1.9	0.4		5500	2014
1400	MANDOVI AT N	GOA	GOA	30	5.5	7.4	81	1.5	0.1		4049	2014
1476	MANDOVI AT T	GOA	GOA	29.2	6.1	6.7	308	1.4	0.3		5672	2014
3185	RIVER MANDO	GOA	GOA	29.6	6.4	6.7	414	1	0.2		9423	2014
3186	RIVER MANDO	GOA	GOA	30	6.4	7.6	305	2.2	0.1		4990	2014
3187	RIVER MANDO	GOA	GOA	30.1	6.3	7.6	77	2.3	0.1		4301	2014
1543	RIVER KALNA A	GOA	GOA	27.8	7.1	7.1	176	1.2	0.1		7817	2014
1548	RIVER ASSONC	GOA	GOA	27.9	6.7	6.4	93	1.4	0.1	2147	3433	2014
2276	RIVER BICHOLII	GOA	GOA	29.3	7.4	6.8	121	1.7	0.4		18125	2014
2275	RIVER CHAPOR	GOA	GOA	29.2	6.9	7	620	1.1	0.1		6300	2014
3189	RIVER CHAPOR	GOA	GOA	30	6	7.5	72	1.6	0.2		9517	2014
1546	RIVER KHANDE	GOA	GOA	29	7.3	7	247	1.5	0.2	1095	2453	2014
2270	RIVER KHANDE	GOA	GOA	29.1	7.3	7	188	1	0.1	1286	3048	2014

- *Excel Code for Cleaning:*

- *State:*

```
=iferror(right(b1046,len(b1046)-find("@",substitute(b1046,",","@",len(b1046)-len(substitute(b1046,",",""))),1)),"")
```

and

```
=trim(clean((substitute(c2,char(160)," "))))
```

- *Data Visualization*

- *Scatter Plot*

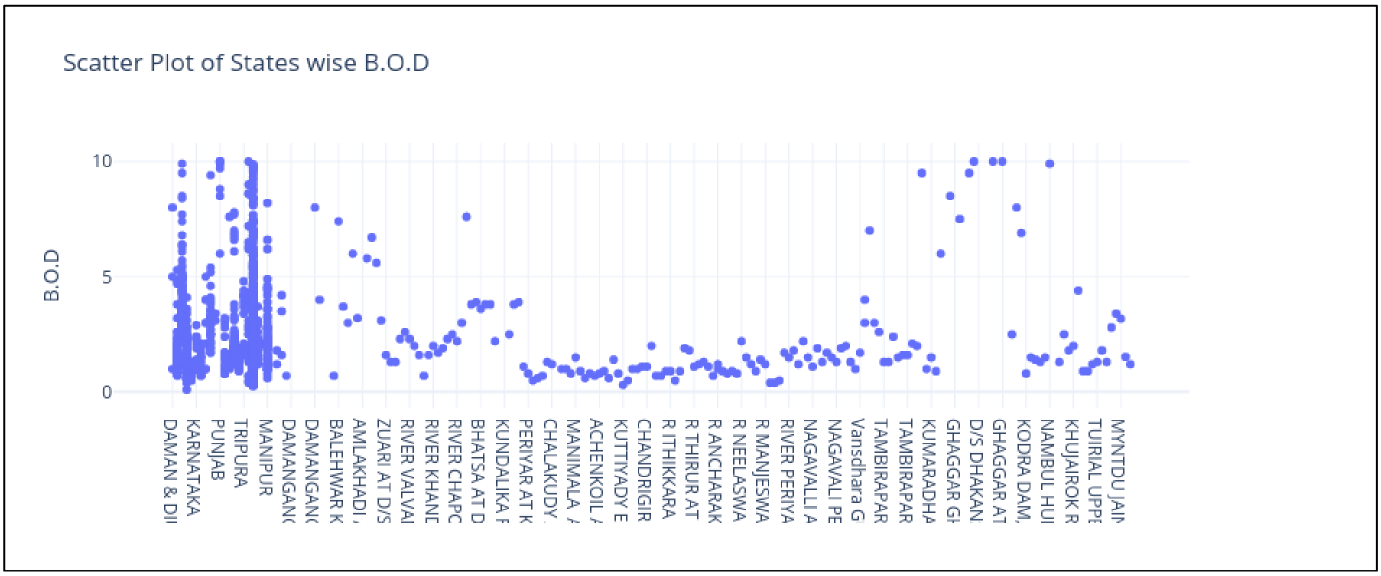


Fig 1: Scatter Plot of State Wise B.O.D

- This scatter plot shows state wise B.O.D
- Here we found maximum B.O.D on Manipur state

- *Second Maximum Found on Daman and Div and Karnataka also* .

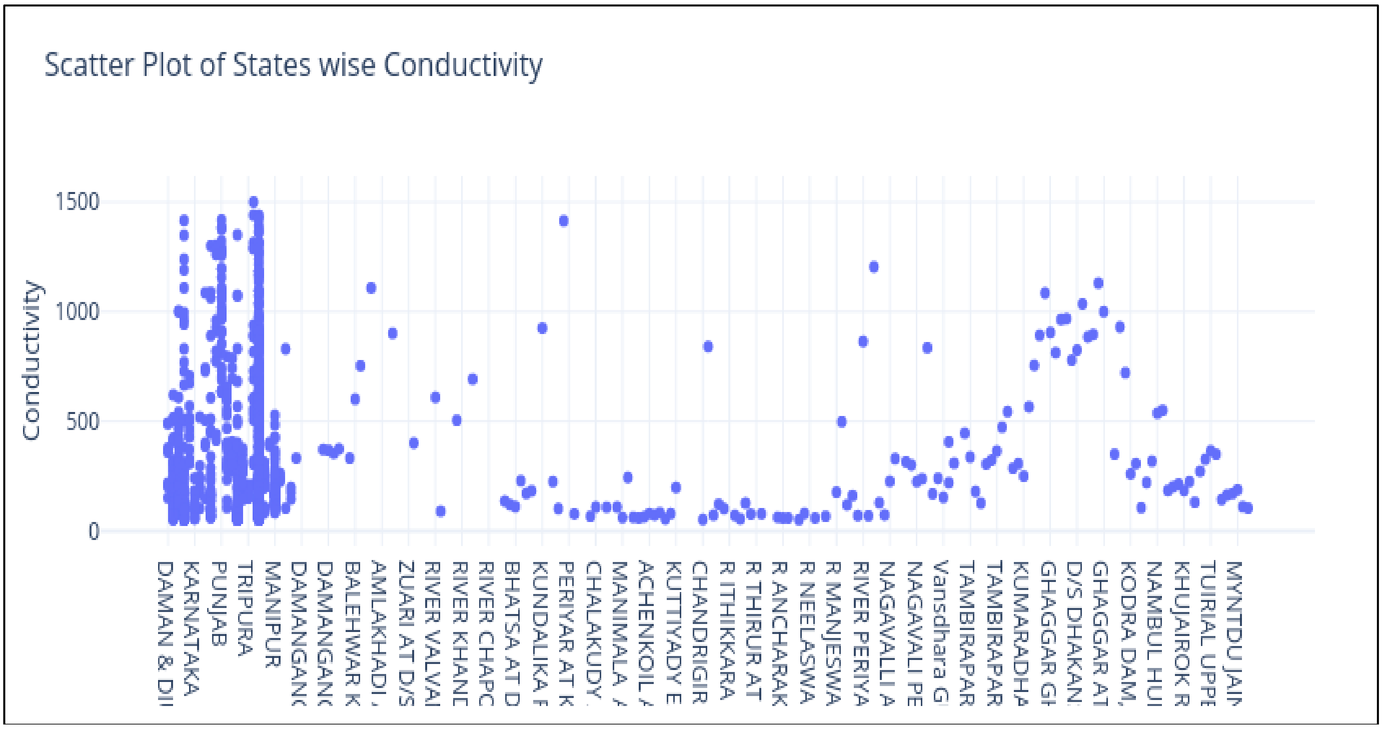


Fig 2: Scatter Plot of State Wise Conductivity

- This scatter plot shows the state wise conductivity
- The maximum conductivity found on Manipur state
- The second maximum conductivity found on Tripura state

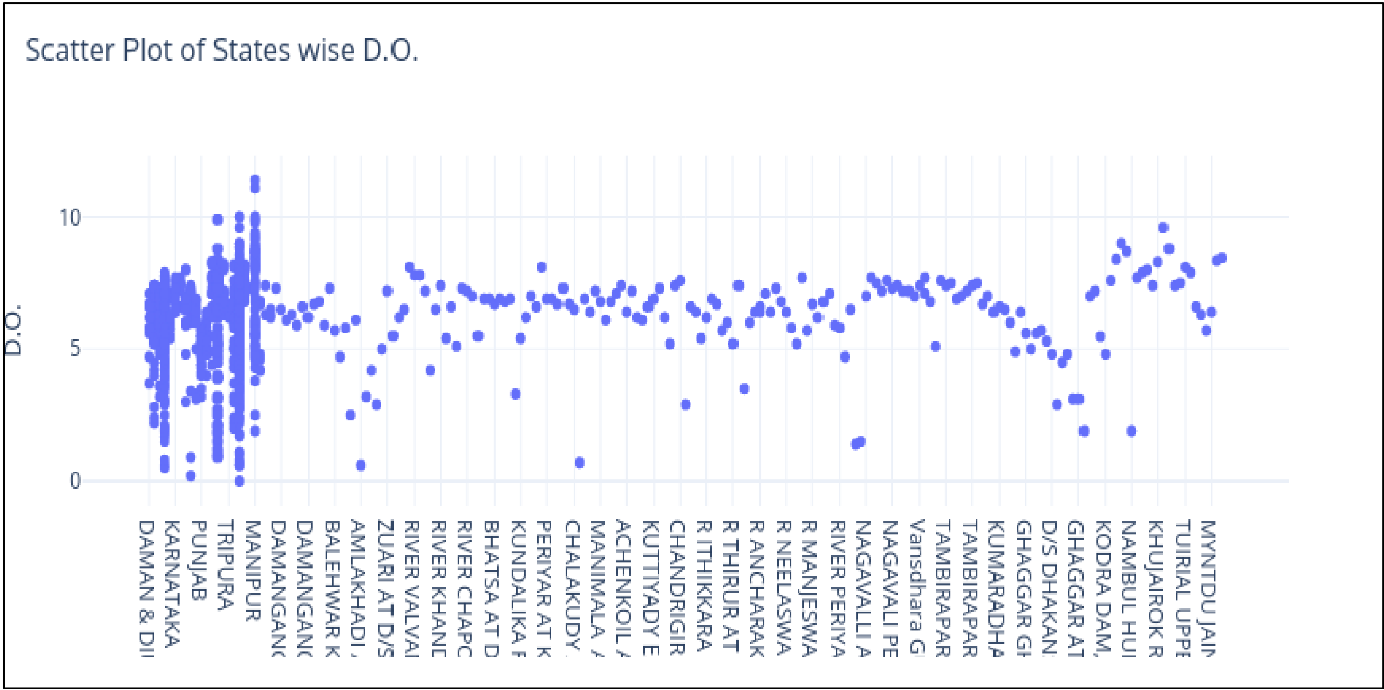


Fig 3: Scatter Plot of State Wise D.O

- This scatter plot shows the state wise D.O
- The maximum D.O can be found on Manipur
- The second maximum D.O can be found on Tripura

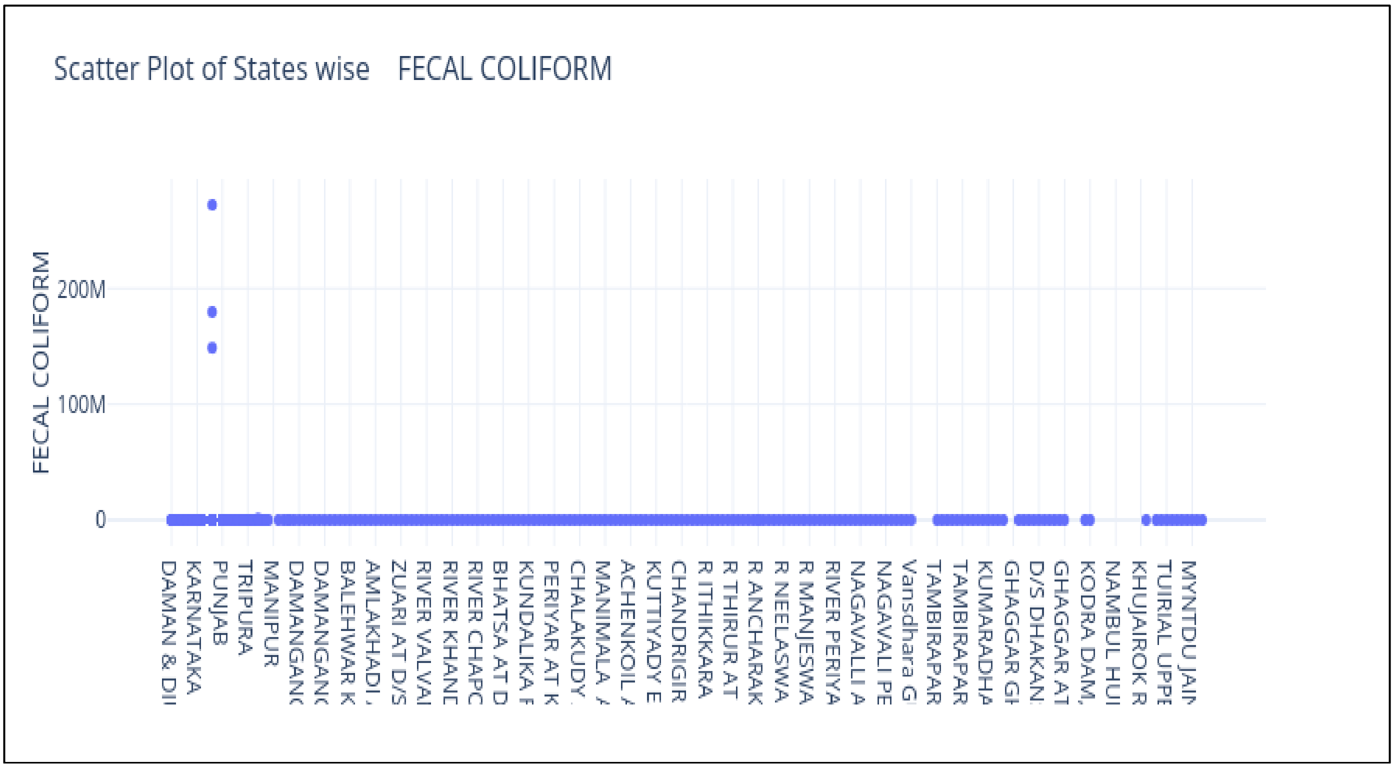


Fig 4: Scatter Plot of State Wise FECAL COLIFORM

- This scatter plot shows the state wise fecal coliform
- Here maximum coliform is not found
- The minimum coliform can be found on Punjab state

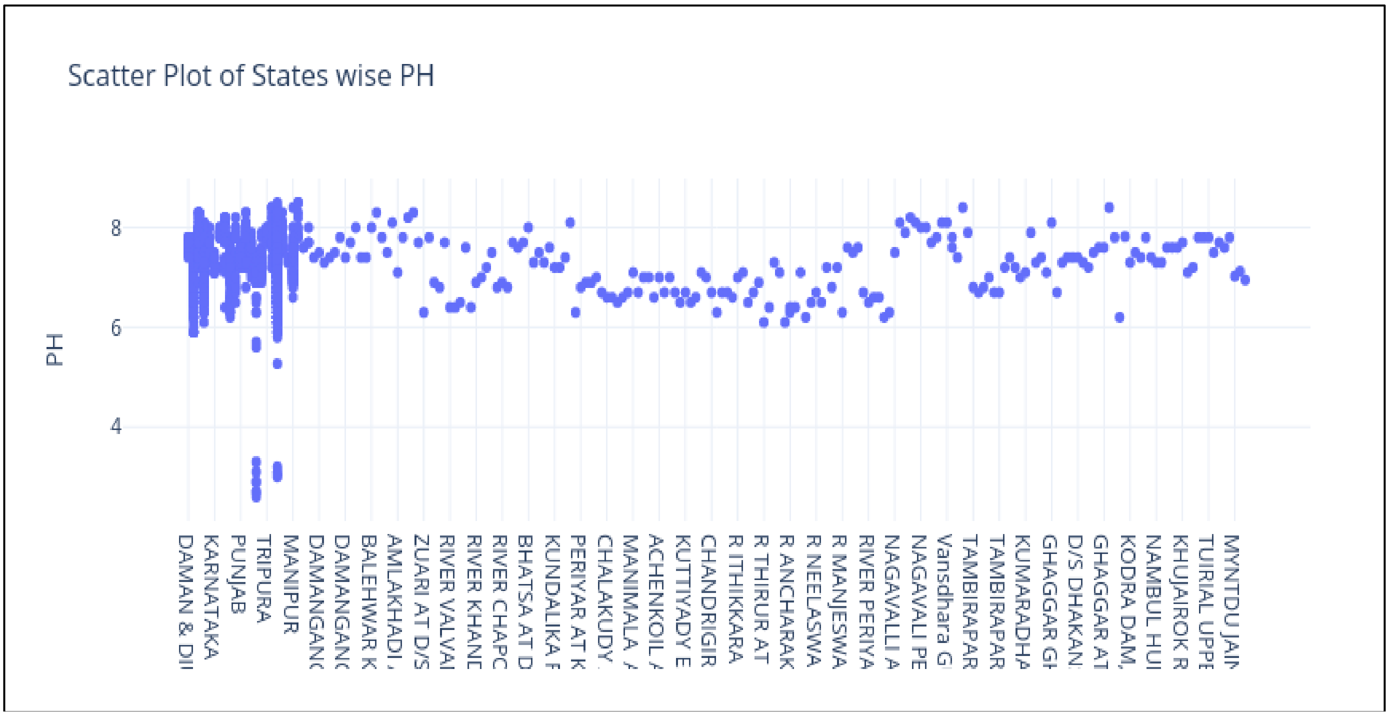


Fig 5: Scatter Plot of State Wise PH

- This scatter plot shows the state wise ph
- Here maximum ph can be found on Manipur state
- The second maximum ph found on Tripura state

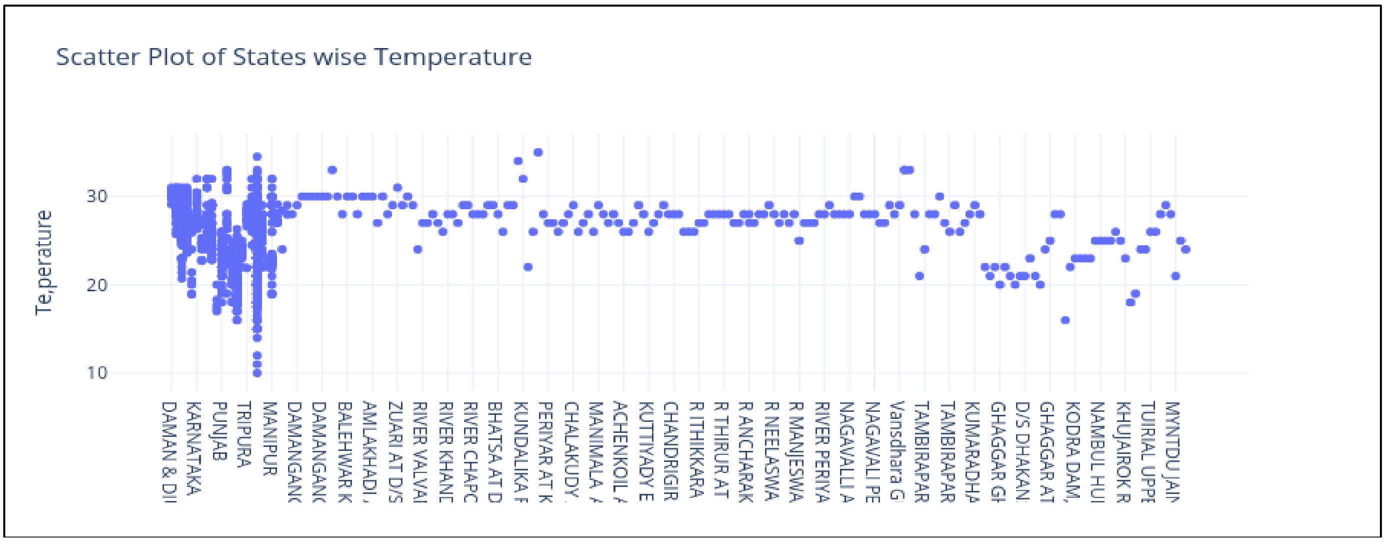


Fig 6: Scatter Plot of State Wise Temperature

- This scatter plot shows state wise temperature
- Here maximum temperature found on Manipur state
- The second maximum temperature found on Tripura state

➢ *Histogram*

```
d<-read.csv("C:\\Users\\swara thakkar\\Desktop\\graph.csv")
d1<-as.vector(d)
 hist(d1$Temp,main = "Histogram of Temprature",col = rainbow(6),xlim = c(0,30))
```

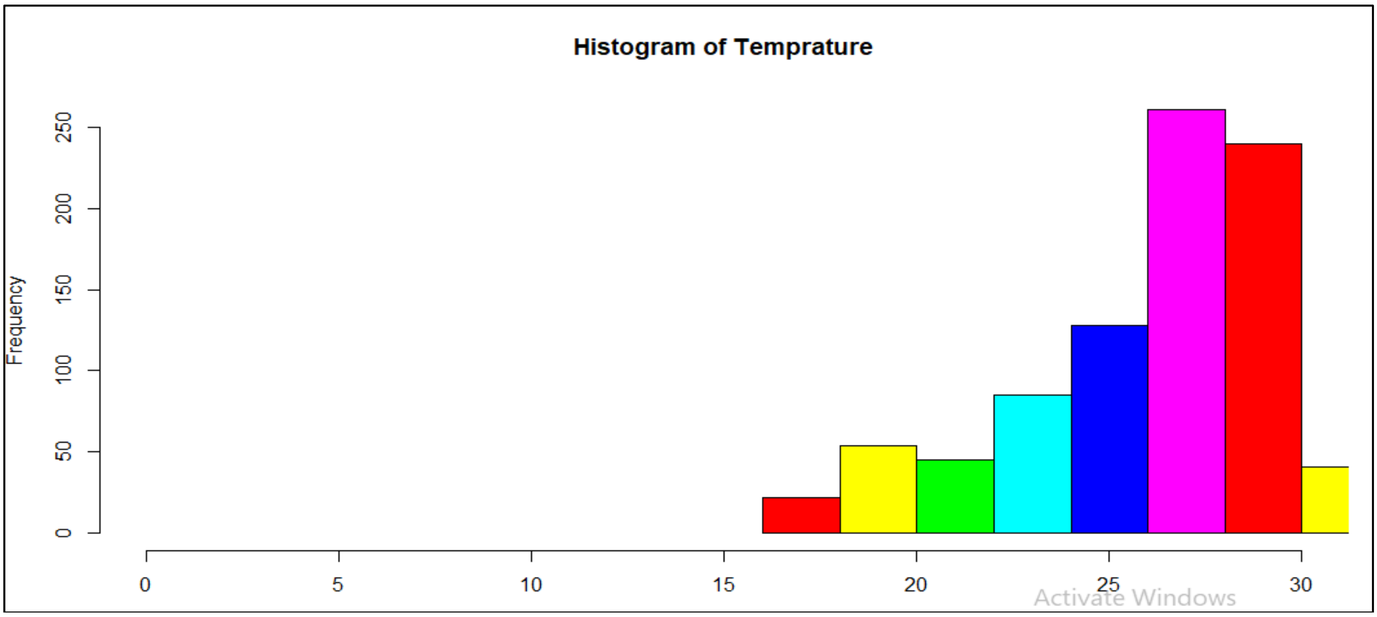


Fig 7: Histogram of Temperature

- This histogram shows the frequency of the temperature

```
d<-read.csv("C:\\Users\\swara thakkar\\Desktop\\graph.csv")
d1<-as.vector(d)
hist(d1$PH,main = "Histogram of P.H",col = rainbow(10),breaks = 10,xlim = c(6,8.5))
```

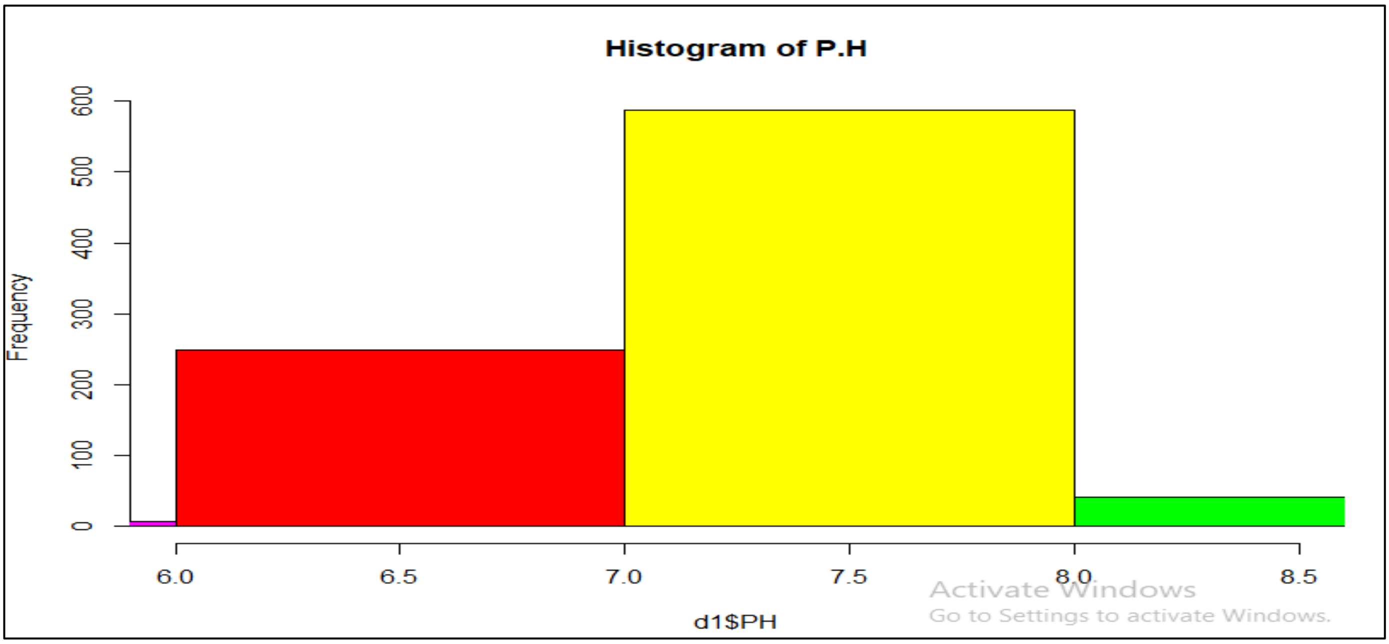


Fig 8: Histogram of P.H

- This histogram shows a frequency of ph
- PH can be increasing when temperature is high

```
d<-read.csv("C:\\Users\\swara thakkar\\Desktop\\graph.csv")
d1<-as.vector(d)
    hist(d1$D.O,main = "Histogram of D.O",col = rainbow(5),breaks = 10,xlim = c(7,14))
```

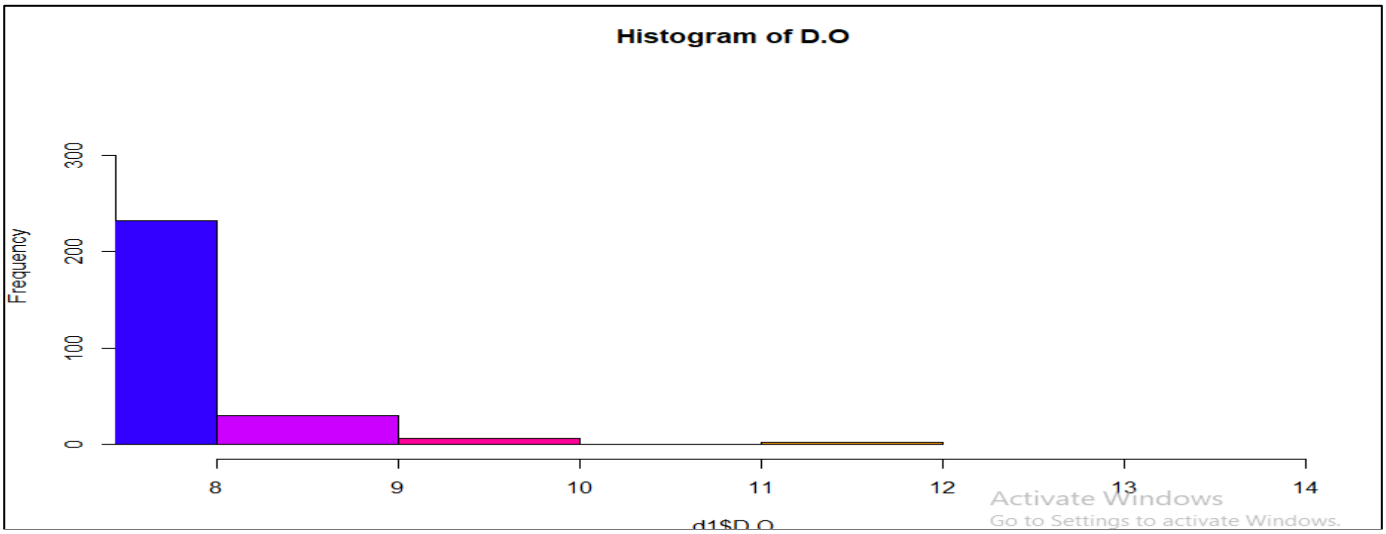


Fig 9: Histogram of D.O

- This histogram shows the frequency of D.O

```
d<-read.csv("C:\\Users\\swara thakkar\\Desktop\\graph.csv")
d1<-as.vector(d)
     hist(d1$B.O.D,main = "Histogram of B.O.D",col = rainbow(10),breaks = 10,xlim = c(2,8))
```

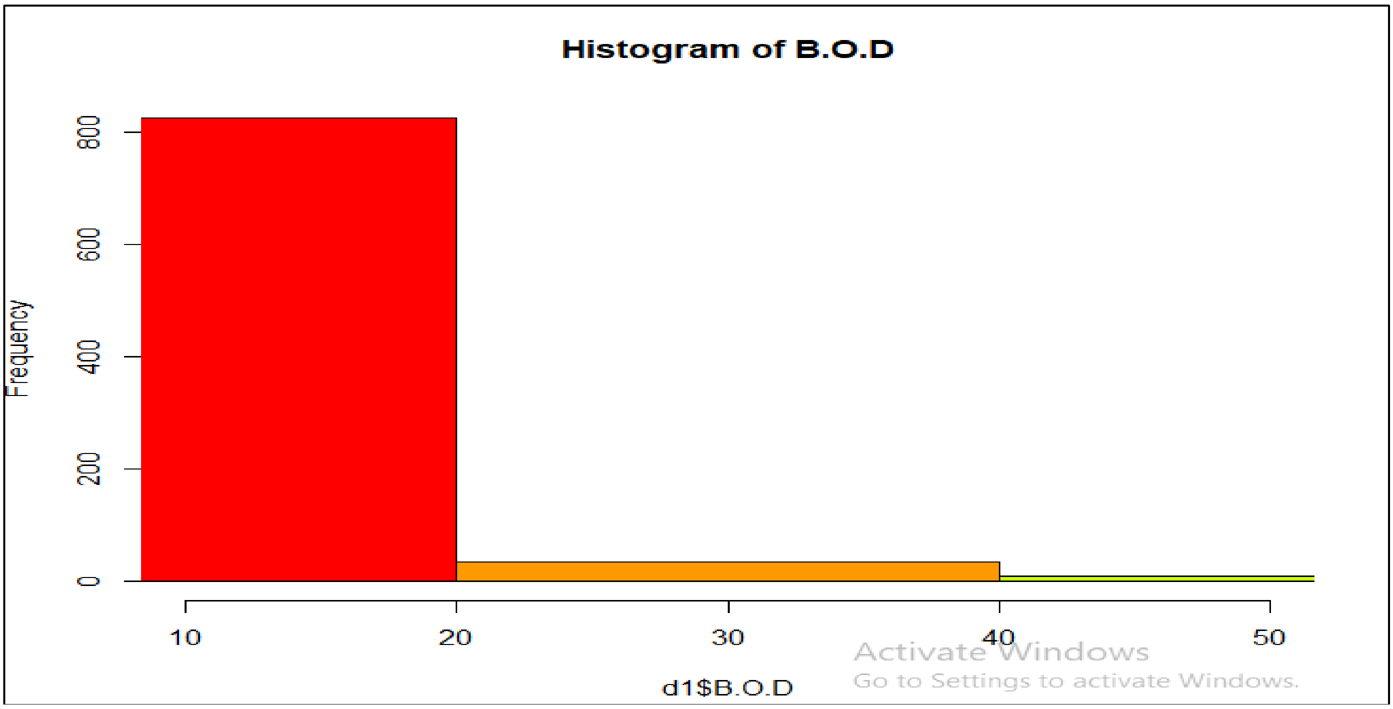


Fig 10: Histogram of B.O.D

- This histogram shows the Frequency of B.O.D

B.O.D value can be lies between 2 to 8

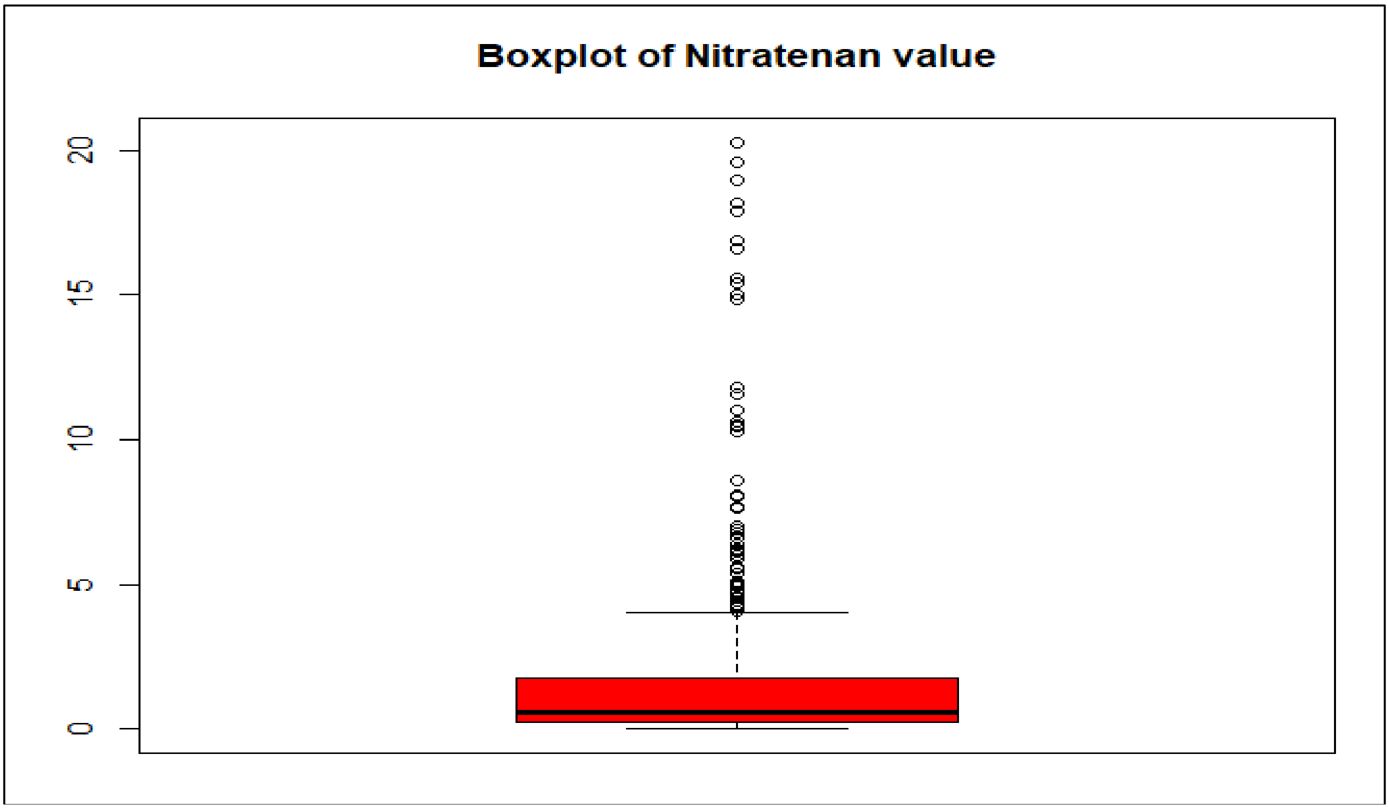


Fig 11: Boxplot of Nitratean Value

- This box plot shows the nitrate value which is going to out of range.

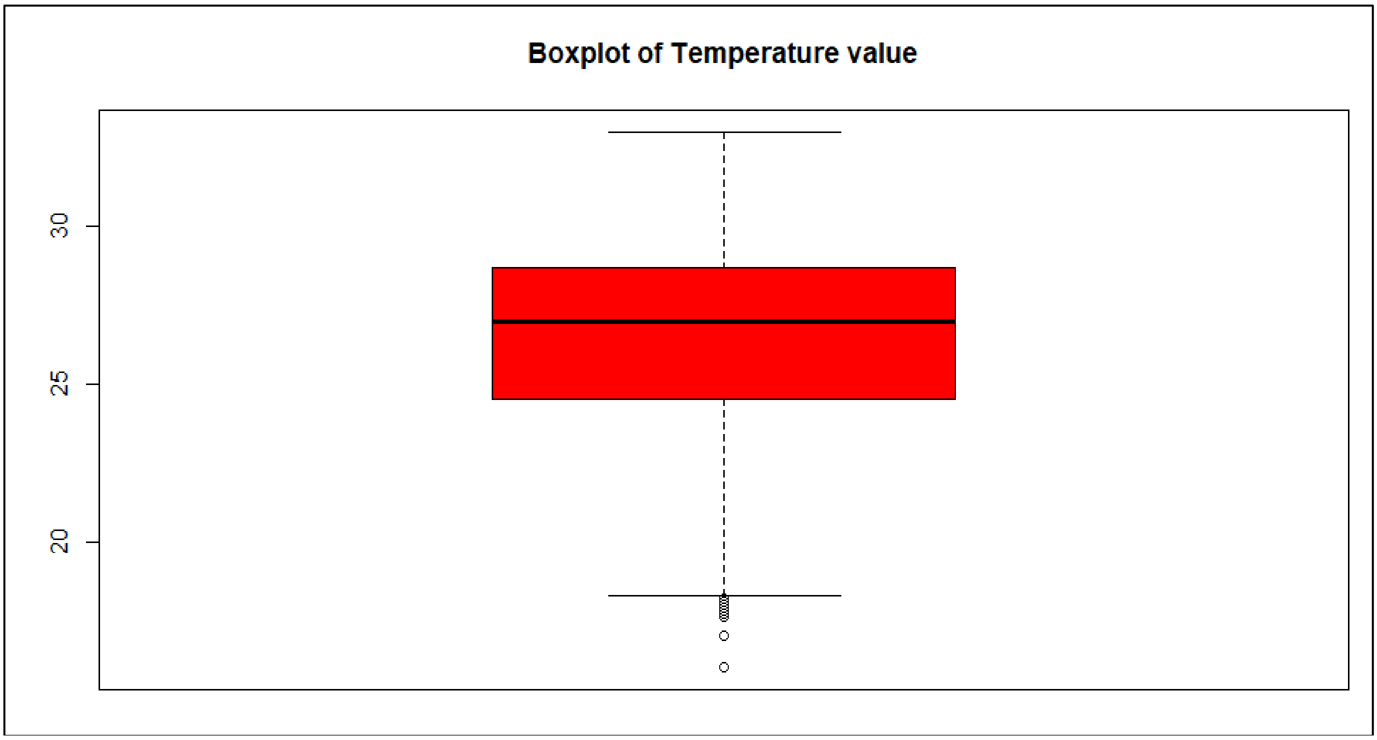


Fig 12: Boxplot of Temperature Value

- This boxplot show the temperature value which is going to out of range.

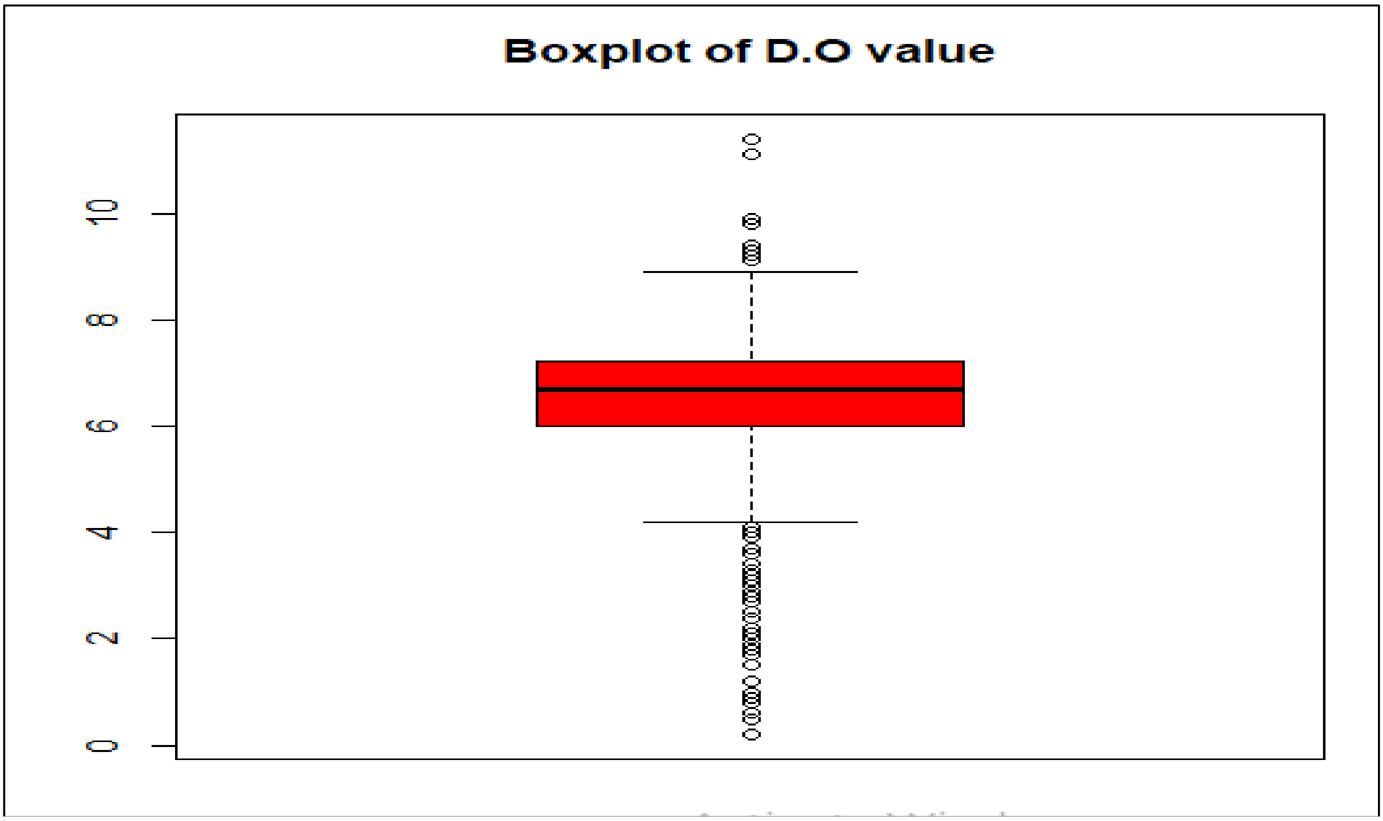


Fig 13: Boxplot of D.O value

- This boxplot shows the D.O value which is going to out of range

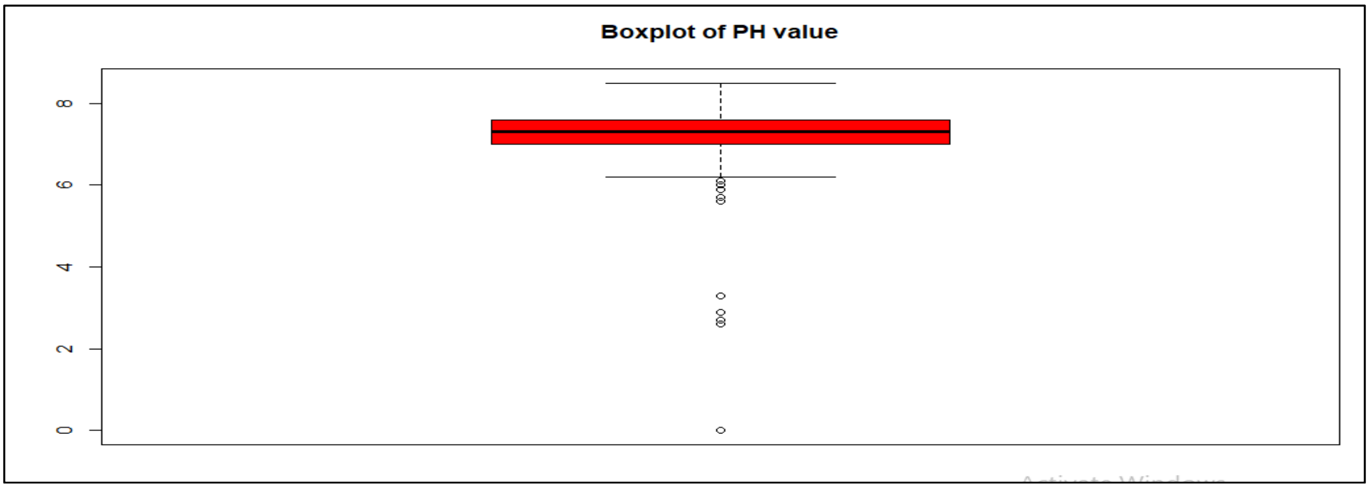
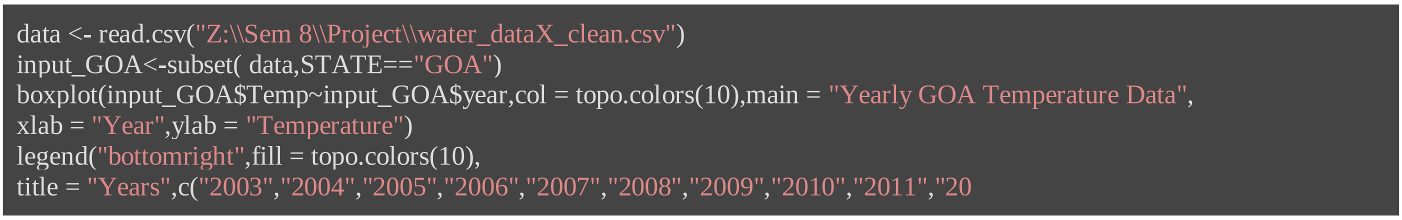


Fig 14: Boxplot of PH value

- This boxplot shows the PH value which is going to out of range



```
data <- read.csv("Z:\\Sem 8\\Project\\water_dataX_clean.csv")
input_GOA<-subset( data,STATE=="GOA")
boxplot(input_GOA$Temp~input_GOA$year,col = topo.colors(10),main = "Yearly GOA Temperature Data",
xlab = "Year",ylab = "Temperature")
legend("bottomright",fill = topo.colors(10),
title = "Years",c("2003","2004","2005","2006","2007","2008","2009","2010","2011","20
```

V. HYPOTHESIS INFORMATION

- There is major co-relation between any two chemical or variables. Such as PH affect the CONDUCTIVITY of water.
- If Temperature increasing or decreasing in GOA, will effects D.O. (mg/l) & PH because of having major part of GOA is connected to water.
- In water quality COLIFORM chemical will be high for next 3 years in kerala state according to previous year's ratio.

➢ *Modeling*

- *Model Identification*

✓ Time series analysis
✓ Linear regression

- *Model Implementation*

✓ Time series analysis

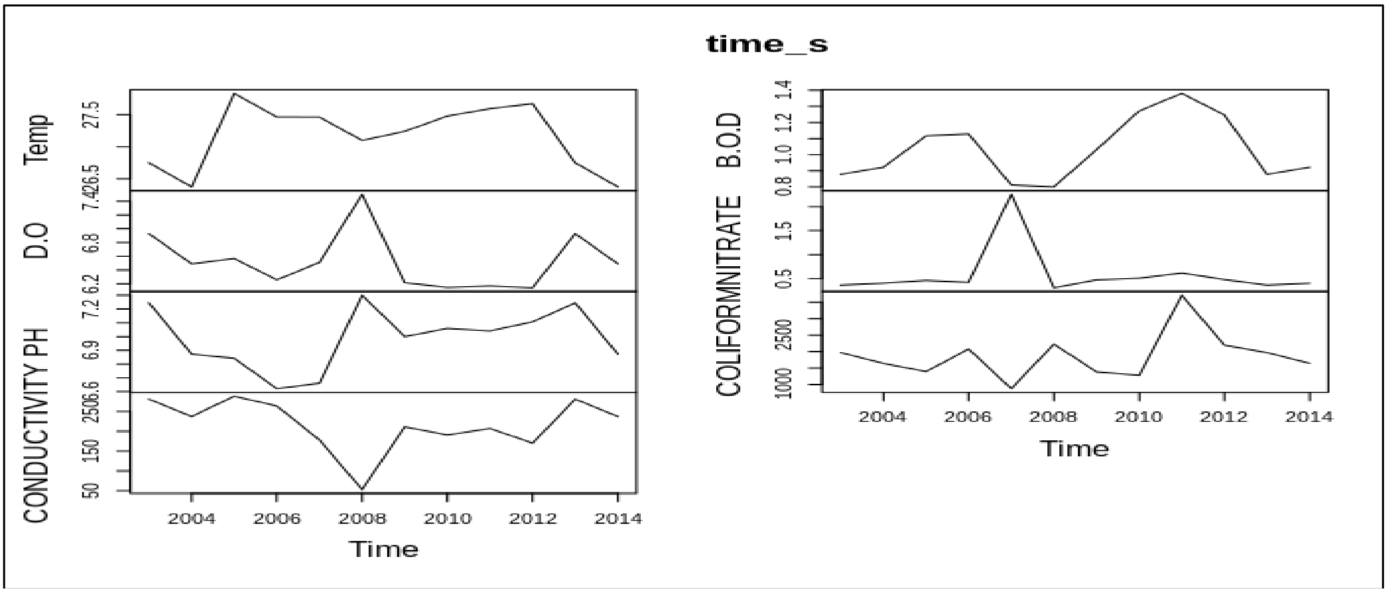


Fig 15: Kerala State

- *Description for Time Series Analysis:*

- *Time Series:*

In Order to perform time series on our data on GOA & KERALA state we performed given steps:

- ✓ We sub-divided (filtered) data according to states
- ✓ Filter the each state data with year (yearly data)
- ✓ For each year starting to end create time series plot.

With the same script we can perform many other time series with different states but here we analysis only most relevant states for analysis.

- *Goa State*

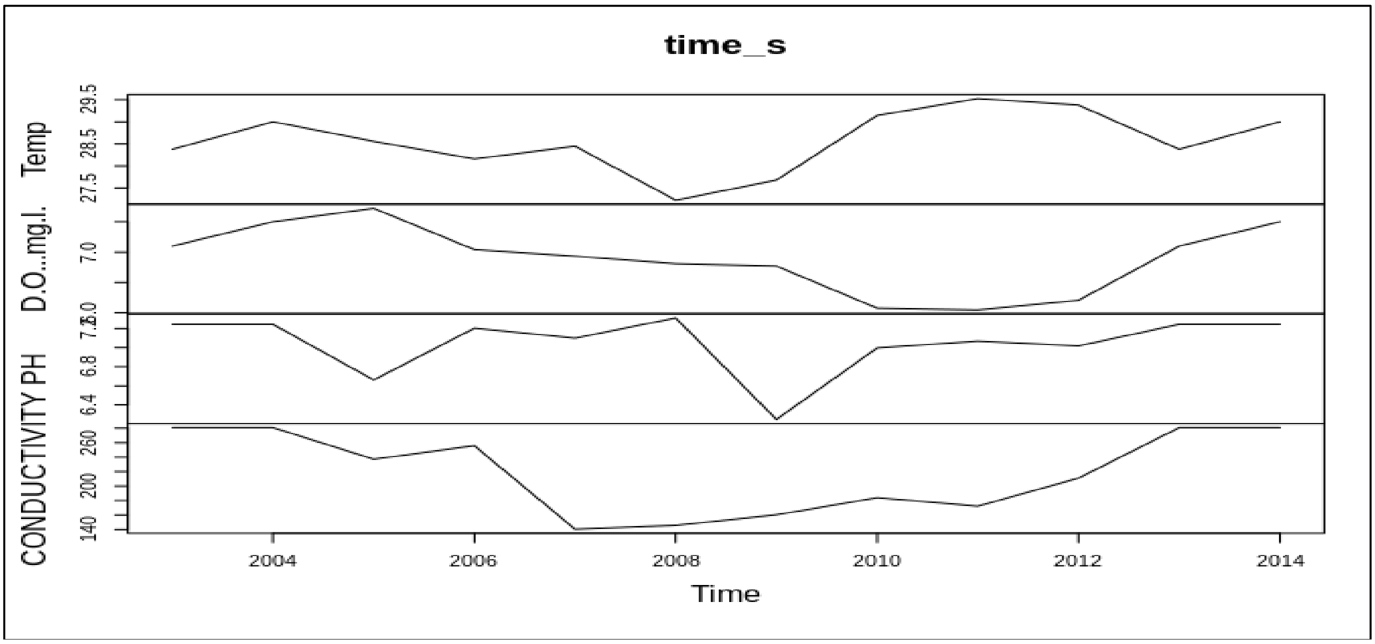


Fig 16: GOA State

For GOA State we can find temperature variance and also conductivity and PH measure are changing on time period.

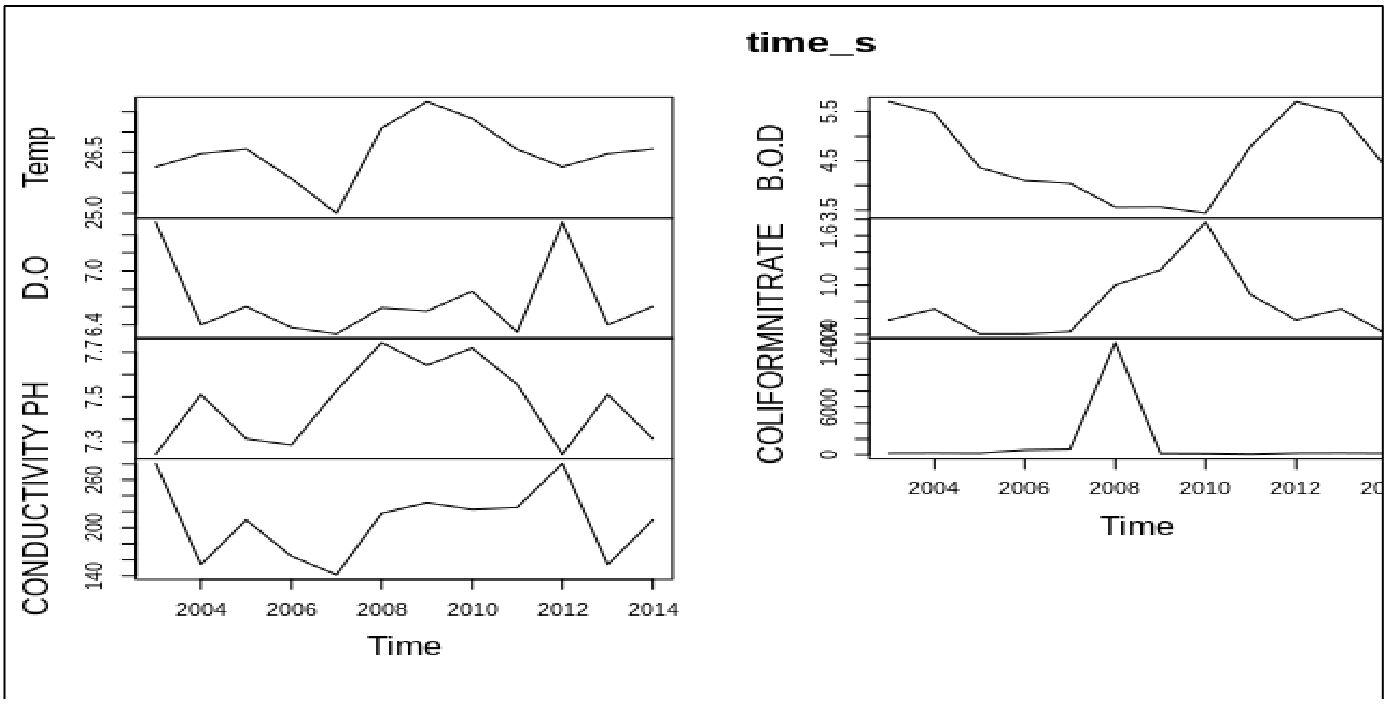
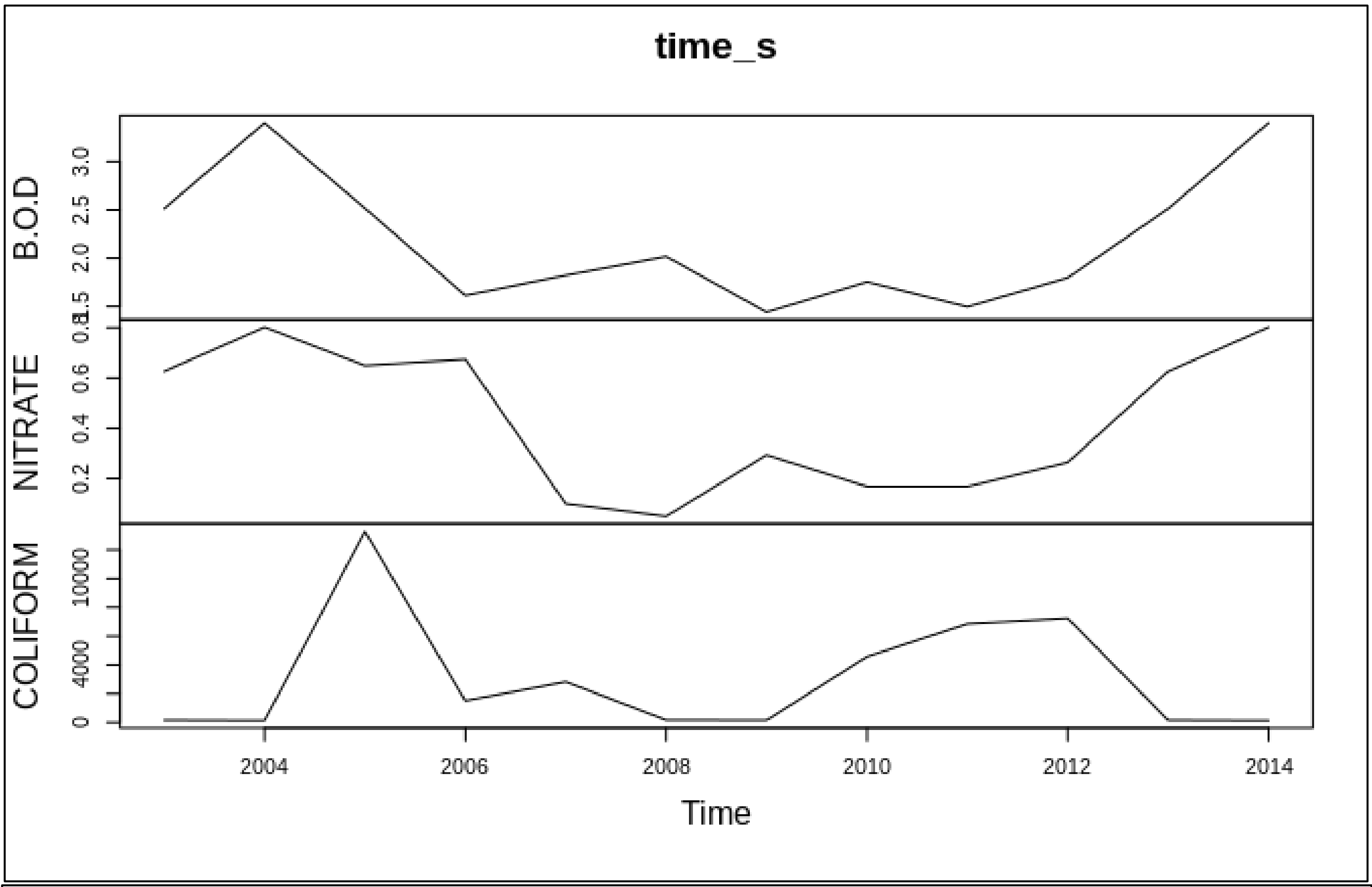


Fig 17: Maharastra State



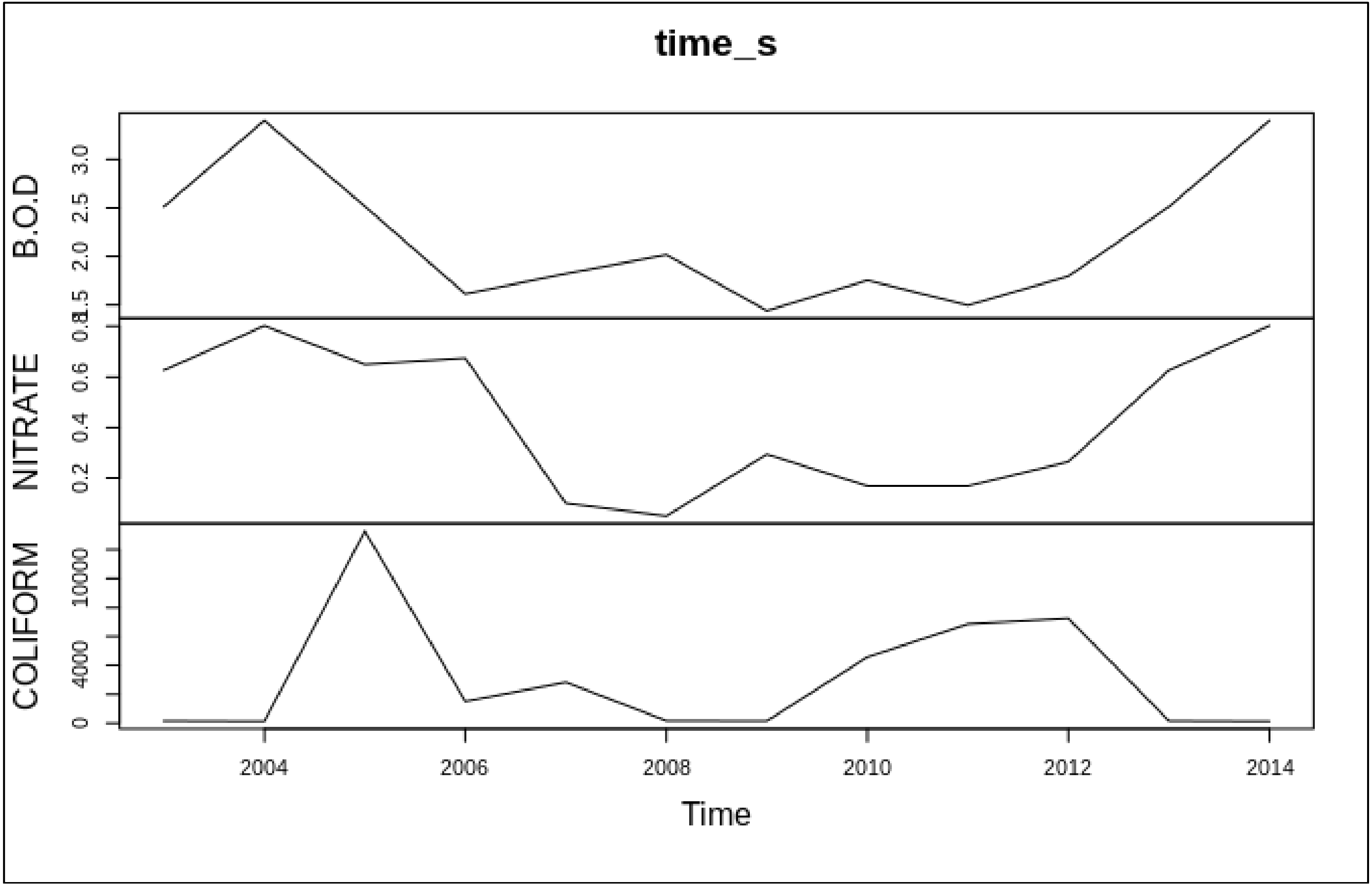
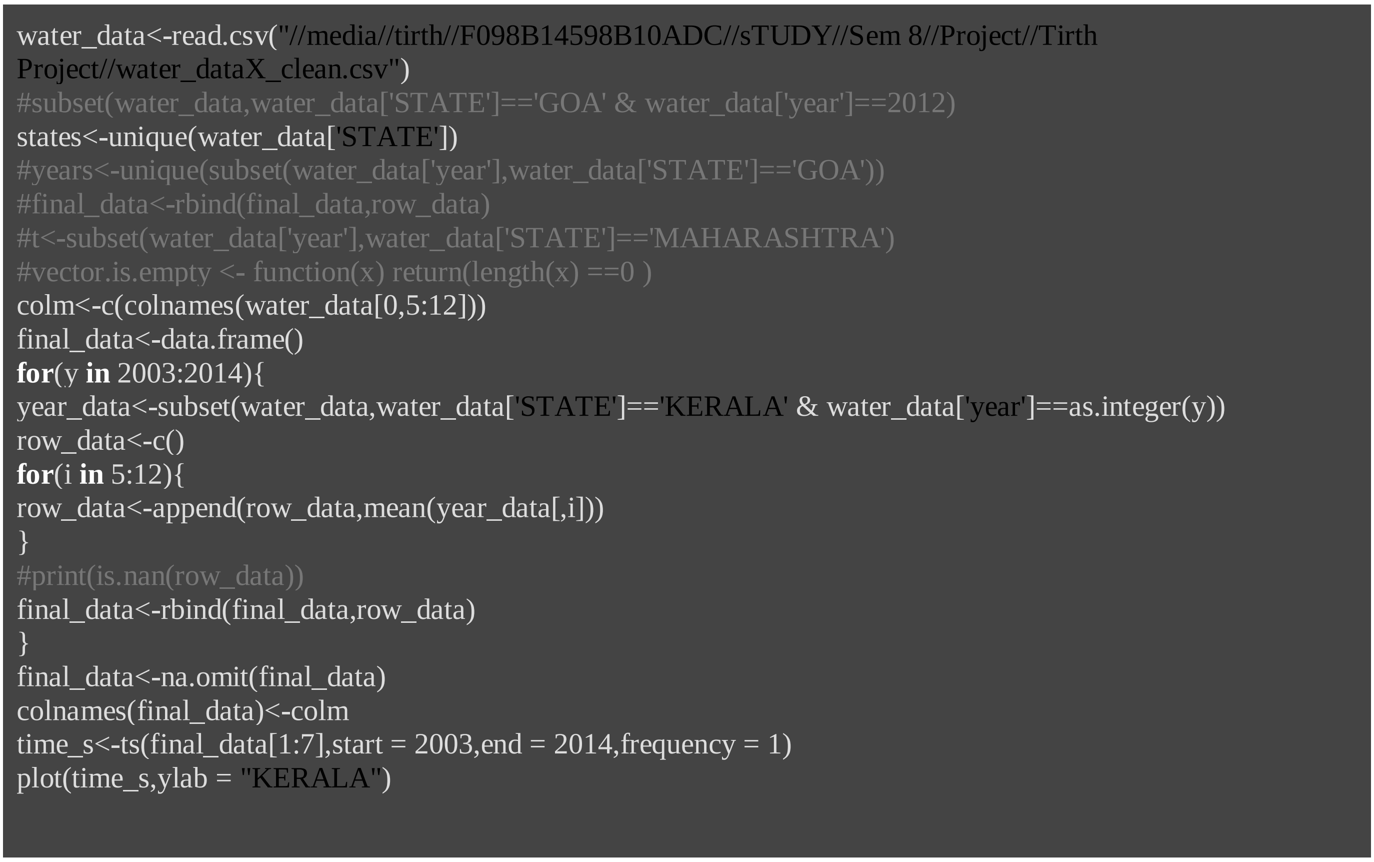


Fig 18: Goa State

VI. CODE FOR TIME SERIES ANALYSIS



```
water_data<-read.csv("//media//tirth//F098B14598B10ADC//sTUDY//Sem 8//Project//Tirth
Project//water_dataX_clean.csv")
#subset(water_data,water_data['STATE']=='GOA' & water_data['year']==2012)
states<-unique(water_data['STATE'])
#years<-unique(subset(water_data['year'],water_data['STATE']=='GOA'))
#final_data<-rbind(final_data,row_data)
#t<-subset(water_data['year'],water_data['STATE']=='MAHARASHTRA')
#vector.is.empty <- function(x) return(length(x) ==0 )
colm<-c(colnames(water_data[0,5:12]))
final_data<-data.frame()
for(y in 2003:2014){
year_data<-subset(water_data,water_data['STATE']=='KERALA' & water_data['year']==as.integer(y))
row_data<-c()
for(i in 5:12){
row_data<-append(row_data,mean(year_data[,i]))
}
#print(is.nan(row_data))
final_data<-rbind(final_data,row_data)
}
final_data<-na.omit(final_data)
colnames(final_data)<-colm
time_s<-ts(final_data[1:7],start = 2003,end = 2014,frequency = 1)
plot(time_s,ylab = "KERALA")
```

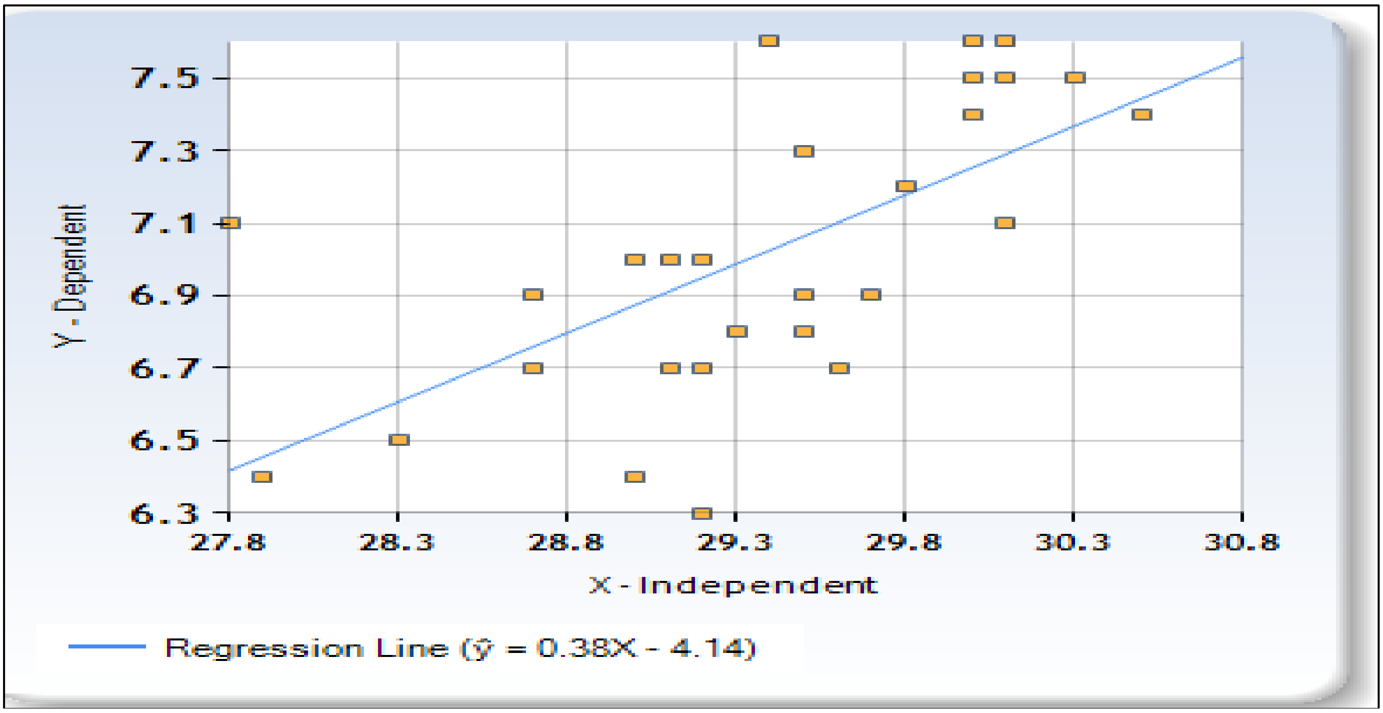


Fig 19: Linear Regression

Here
X axis shows the temperature value
Y axis shows the ph value
If temperature will increase ph will also increase

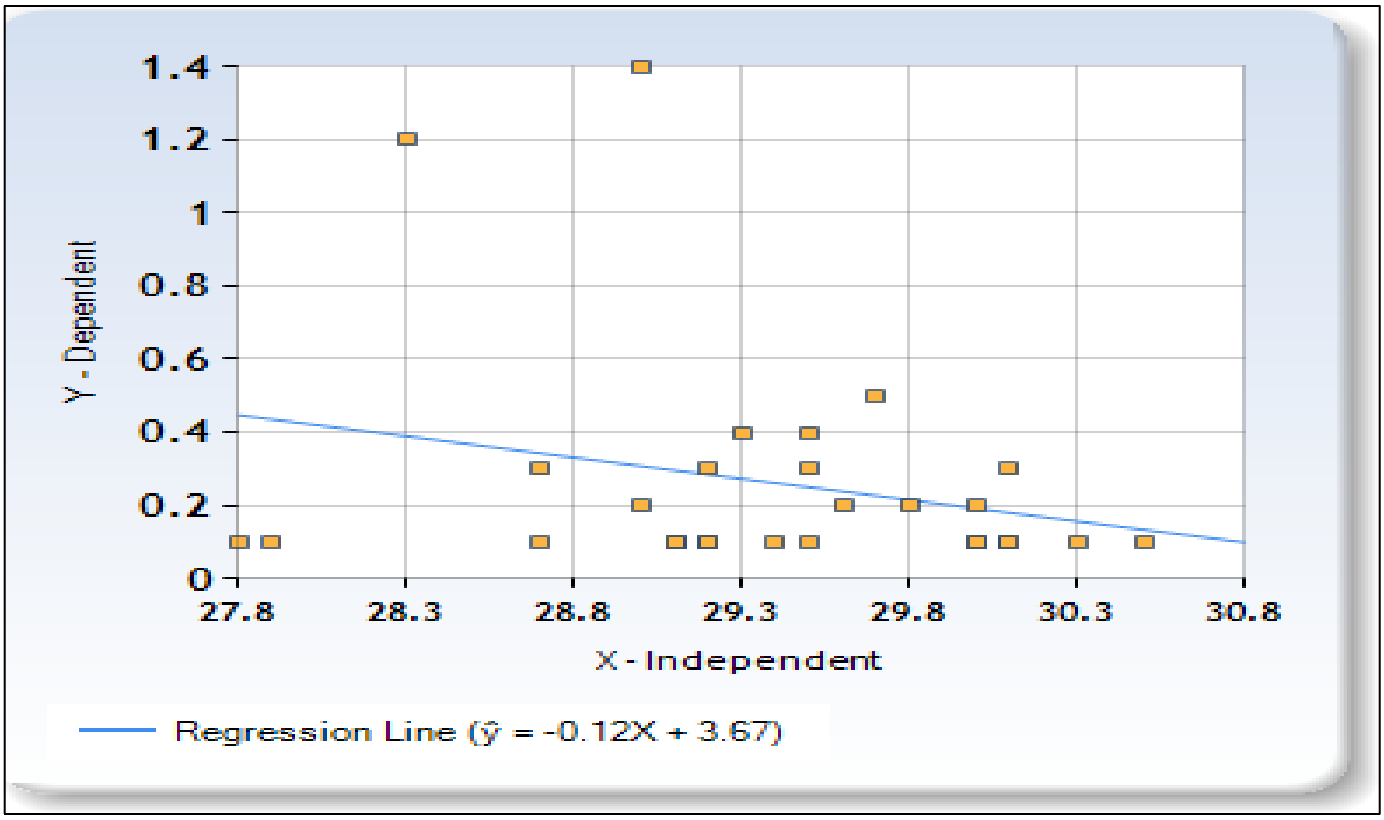


Fig 20: Regressions

HERE
X shows the temperature value
Y axis shows the nitrate value
IF temperature is increase nitrate will decrease

VII. DISCUSSION AND CONCLUSION

- We analyzed a dataset that contains information about important parameter or chemical of water. The dataset was collected from 2003 to 2014. It contains 1992 records.
- Maximum B.O.D, D.O and conductivity values can be finding on Manipur state.
- PH and nitrate depended on temperature. If temperature is increase nitrate will decrease
- Based on time series analysis we can say that B.O.D value can be increase in 2004 year

FUTURE DIRECTION

- By analyzing, we get to know on which state water quality is good.
- We also get to know about which chemical is more important for water quality?
- Which factor is affected to water quality?

REFERENCES

➢ *Web Link:*

[1]. https://www.kaggle.com/venkatramakrishnan/india-water-quality-data

[2]. https://www.geeksforgeeks.org/

[3]. https://en.wikipedia.org/wiki/Water_quality

[4]. https://www.intechopen.com/chapters/69568

[5]. https://www.cgwb.gov.in/old_website/wqreports.html

[6]. https://en.wikipedia.org/wiki/Water_pollution_in_India

[7]. https://www.fairplanet.org/story/india-how-water-pollution-triggers-migration-waves/

➢ *Book References:*

[8]. Thomas Erl,Wajid Khattak,and Paul Buhler: Big Data Fundamentals: Concepts, Dribers and techniques , Pearson.