Project on water potability
Explanatory Analysis
Presidency University

Shubhamoy Paul

July 17, 2022

I am took a water potability dataset from kaggle.Data file name is **water_potability.csv**

Contex

Access to safe drinking-water is essential to health, a basic human right and a component of effective policy for health protection. This is important as a health and development issue at a national, regional and local level. In some regions, it has been shown that investments in water supply and sanitation can yield a net economic benefit, since the reductions in adverse health effects and health care costs outweigh the costs of undertaking the interventions.

Content

The water_potability.csv file contains water quality metrics for 3276 different water bodies.

pH value: PH is an important parameter in evaluating the acid–base balance of water. It is also the indicator of acidic or alkaline condition of water status. WHO has recommended maximum permissible limit of pH from 6.5 to 8.5. The current investigation ranges were 6.52–6.83 which are in the range of WHO standards.

Hardness: Hardness is mainly caused by calcium and magnesium salts. These salts are dissolved from geologic deposits through which water travels. The length of time water is in contact with hardness producing material helps determine how much hardness there is in raw water. Hardness was originally defined as the capacity of water to precipitate soap caused by Calcium and Magnesium.

Solids (Total dissolved solids - TDS): Water has the ability to dissolve a wide range of inorganic and some organic minerals or salts such as potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates etc. These minerals produced un-wanted taste and diluted color in appearance of water. This is the important parameter for the use of water. The water with high TDS value indicates that water is highly mineralized. Desirable limit for TDS is 500 mg/l and maximum limit is 1000 mg/l which prescribed for drinking purpose.

Chloramines: Chlorine and chloramine are the major disinfectants used in public water systems. Chloramines are most commonly formed when ammonia is added to chlorine to treat drinking water. Chlorine levels up to 4 milligrams per liter (mg/L or 4 parts per million (ppm)) are considered safe in drinking water.

Sulfate: Sulfates are naturally occurring substances that are found in minerals, soil, and rocks. They are present in ambient air, groundwater, plants, and food. The principal commercial use of sulfate is in the chemical industry. Sulfate concentration in seawater is about 2,700 milligrams per liter (mg/L).

It ranges from 3 to 30 mg/L in most freshwater supplies, although much higher concentrations (1000 mg/L) are found in some geographic locations.

Conductivity: Pure water is not a good conductor of electric current rather's a good insulator. Increase in ions concentration enhances the electrical conductivity of water. Generally, the amount of dissolved solids in water determines the electrical conductivity. Electrical conductivity (EC) actually measures the ionic process of a solution that enables it to transmit current. According to WHO standards, EC value should not exceeded 400 μS/cm.

Organic_carbon: Total Organic Carbon (TOC) in source waters comes from decaying natural organic matter (NOM) as well as synthetic sources. TOC is a measure of the total amount of carbon in organic compounds in pure water. According to US EPA $<$ 2 mg/L as TOC in treated / drinking water, and $<$ 4 mg/Lit in source water which is use for treatment.

Trihalomethanes: THMs are chemicals which may be found in water treated with chlorine. The concentration of THMs in drinking water varies according to the level of organic material in the water, the amount of chlorine required to treat the water, and the temperature of the water that is being treated. THM levels up to 80 ppm is considered safe in drinking water.

Turbidity: The turbidity of water depends on the quantity of solid matter present in the suspended state. It is a measure of light emitting properties of water and the test is used to indicate the quality of waste discharge with respect to colloidal matter. The mean turbidity value obtained for Wondo Genet Campus (0.98 NTU) is lower than the WHO recommended value of 5.00 NTU.

Potability: Indicates if water is safe for human consumption where 1 means Potable and 0 means Not potable.

We are doing explonetory analysis on it then we will fit a model on it.
Some basic step we have to follow for analyzing a dataset.this

step 2. Cleaning the dataset

step 1. Loading the dataset

step 3. Visualized the dataset

step 4. Fit a model on it

Step 1:

I load water portability data from my local file.and show 1st nth row of the dataset using **head()**

```
options(warn=-1)
```

```
data =read.csv("C:/Users/SVMY/Downloads/R project/water_potability.csv")
head(data)
```

```
        ph Hardness   Solids Chloramines  Sulfate Conductivity Organic_carbon
1       NA 204.8905 20791.32    7.300212 368.5164     564.3087      10.379783
2 3.716080 129.4229 18630.06    6.635246       NA     592.8854      15.180013
3 8.099124 224.2363 19909.54    9.275884       NA     418.6062      16.868637
4 8.316766 214.3734 22018.42    8.059332 356.8861     363.2665      18.436524
5 9.092223 181.1015 17978.99    6.546600 310.1357     398.4108      11.558279
6 5.584087 188.3133 28748.69    7.544869 326.6784     280.4679       8.399735
  Trihalomethanes Turbidity Potability
1        86.99097  2.963135          0
2        56.32908  4.500656          0
```

```
3        66.42009  3.055934          0
4       100.34167  4.628771          0
5        31.99799  4.075075          0
6        54.91786  2.559708          0
```

I want to see columm names and dimmension of our dataset.

```
colnames(data) #columms name

 [1] "ph"              "Hardness"        "Solids"          "Chloramines"
 [5] "Sulfate"         "Conductivity"    "Organic_carbon"  "Trihalomethanes"
 [9] "Turbidity"       "Potability"

dim(data)   #dimmension of data

[1] 3276   10
```

Our dataset contains 3276 rows and 10 columms.

Step 2:

we want to is there any null value in dataset.

```
sum(is.na(data))  #sum of null values if in dataset

[1] 1434
```

It shows that there 1434 cells are empty.Now we have to remove these empty from our dataset otherwise it will make trouble to analysis our dataset.

```
data=na.omit(data)
```

```
sum(is.na(data))

[1] 0
```

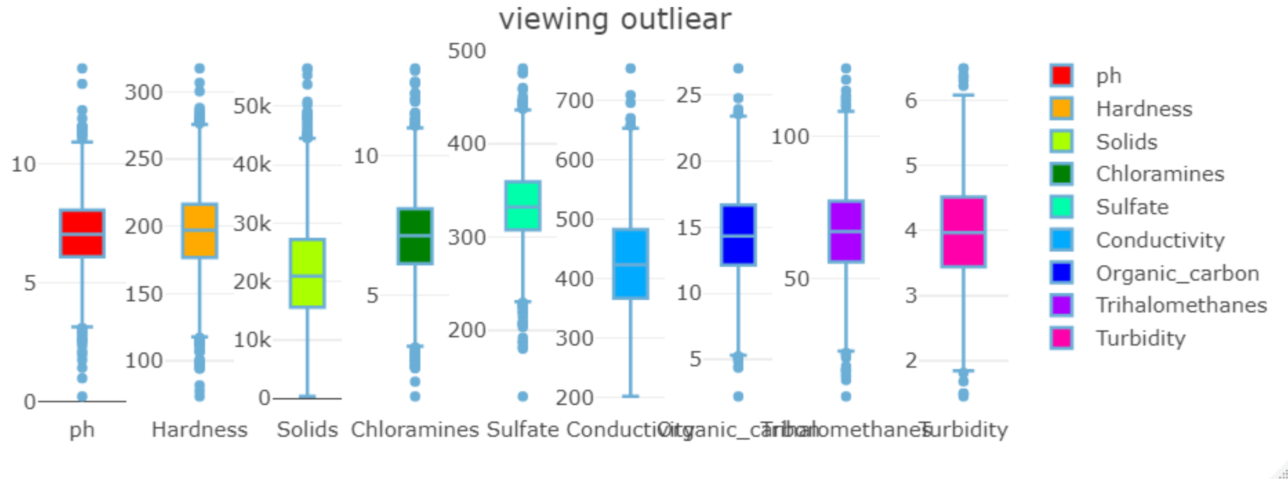
we are interest to type of each variables.

```
sapply(data, typeof)

             ph        Hardness          Solids     Chloramines         Sulfate
       "double"        "double"        "double"        "double"        "double"
   Conductivity  Organic_carbon Trihalomethanes       Turbidity      Potability
       "double"        "double"        "double"        "double"       "integer"
```

here data types are integer and double(float) .so we no need to change any thing.
Outlier

```
options(warn=-1)
library(plotly)
library(webshot)
l <- htmltools::tagList()
for(i in 1:9){
l[[i]]=as.widget(plot_ly(y = data[,i], boxpoints = 'outliers', type = "box",name = colnames(data)[i],mar
        line = list(color = 'rgb(107,174,214)'),fillcolor = palette(rainbow(9))[i]))
```

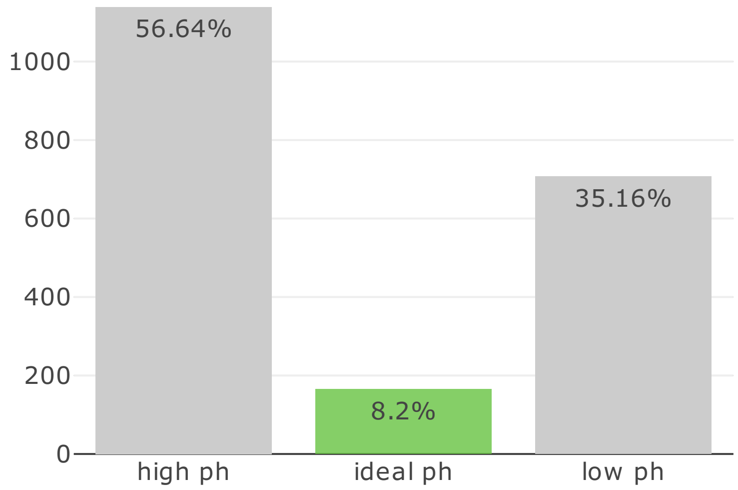
```
}

fig <- subplot(l[[1]], l[[2]],l[[3]],l[[4]],l[[5]],l[[6]],l[[7]],l[[8]],l[[9]]) %>% layout(title = 'view

fig
```



Now visulizing each variables which define water purity.

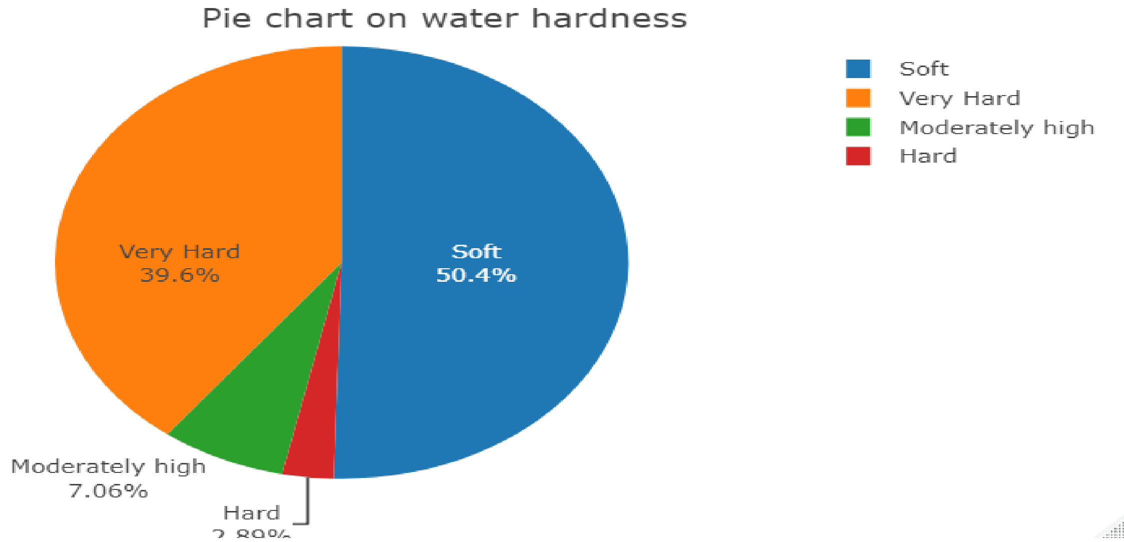
pH value:

```
library("webshot")
library(plotly)

FALSE Loading required package:  ggplot2
FALSE
FALSE Attaching package:  'plotly'
FALSE The following object is masked from 'package:ggplot2':
FALSE
FALSE     last_plot
FALSE The following object is masked from 'package:stats':
FALSE
FALSE     filter
FALSE The following object is masked from 'package:graphics':
FALSE
FALSE     layout

x=c(length(which(data$ph<6.52)),length(which(data$ph>6.52 & data$ph<6.83)),length(which(data$ph>6.83)))
txt=c(paste0(round((x[1]/sum(x))*100,2), "%"),paste0(round((x[2]/sum(x))*100,2), "%"),paste0(round((x[3
fig <- plot_ly(
  x = c("low ph", "ideal ph", "high ph"),
  y = x,
  name = "SF Zoo",
  type = "bar",
  text=txt,
  marker = list(color = c('rgba(204,204,204,1)', 'rgba(103,195,66,0.8)',
                          'rgba(204,204,204,1)'))
)

fig
```



Hardness:

```
label=c("Soft","Moderately high","Hard","Very Hard")
values=c(length(which(data>0 & data<60)),length(which(data>61 & data<120)),length(which(data>121 & data<
fig <- plot_ly(type='pie', labels=label, values=values,
               textinfo='label+percent',
               insidetextorientation='radial')
fig <- fig %>% layout(title = "Pie chart on water hardness")
fig
```



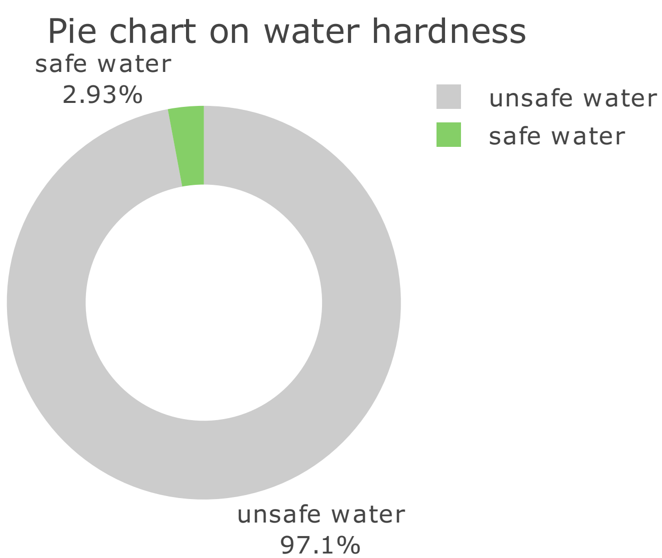
Solids:

there is only one value which is below 1000 this means that all the water is safe coresponding this solids variable.

```
length(which(data$Solids<1000))

[1] 1
```

Chloramines:

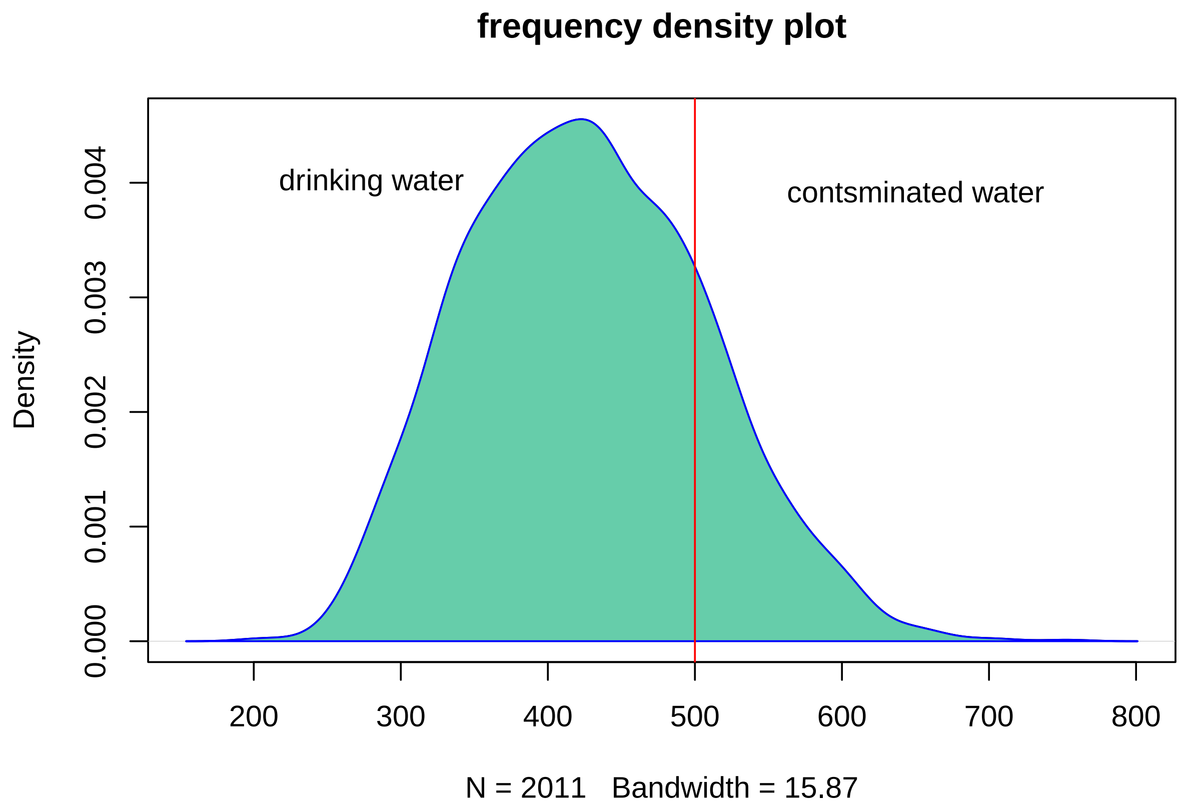
```
label=c("safe water","unsafe water")
values=c(length(which( data$Chloramines<4)),length(which(data$Chloramines>4)))
colors <- c('rgba(103,195,66,0.8)','rgba(204,204,204,1)')
fig <- plot_ly(, labels=label, values=values,
               textinfo='label+percent',marker = list(colors = colors))
fig <- fig %>% add_pie(hole = 0.6)
fig <- fig %>% layout(title = "Pie chart on water hardness")

fig
```



Sulfate:

```
plot(density(data$Conductivity), main="frequency density plot")
polygon(density(data$Conductivity), col="aquamarine3", border="blue")
abline(v=500,col="red")
```

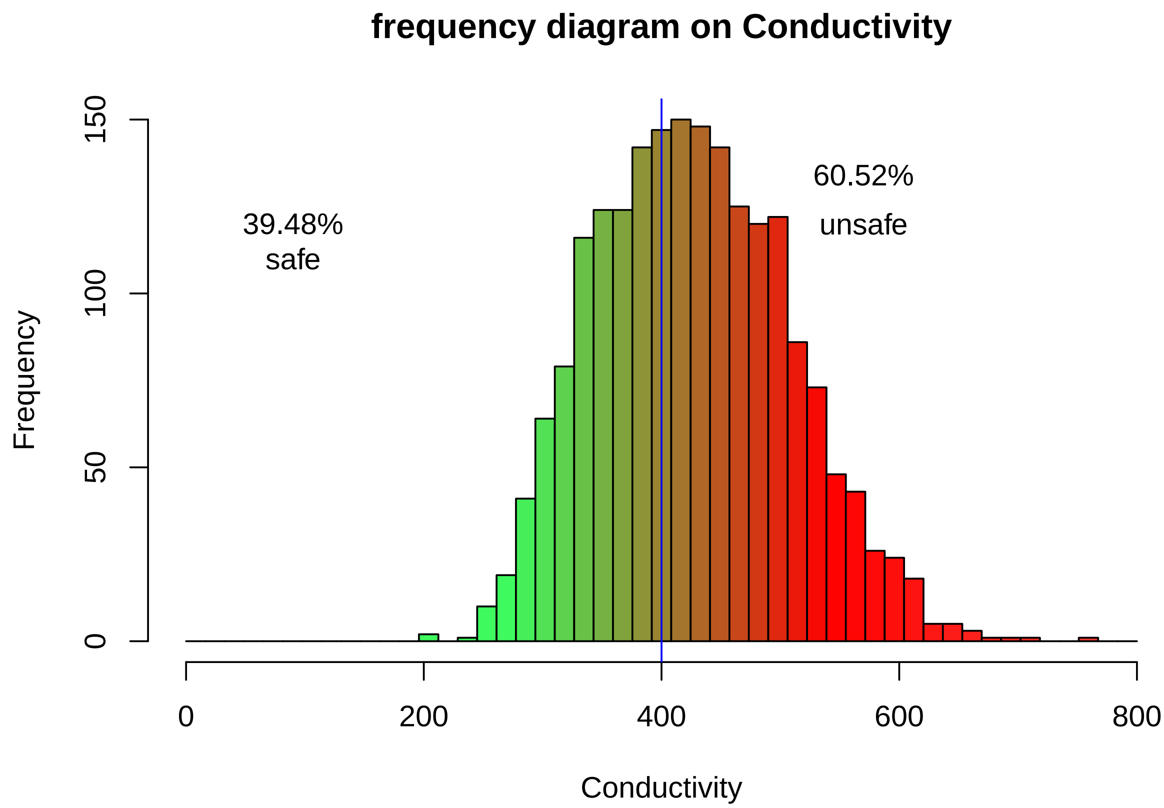
```
#locator() #this function is use to locate point in the diagram
text(280, 0.003999010, expression("drinking water"))
text(650.0549, 0.00392039, expression("contsminated water"))
```



Area right side of red line is dirty water,we can't drink this water & area left side of red line is drinking water.

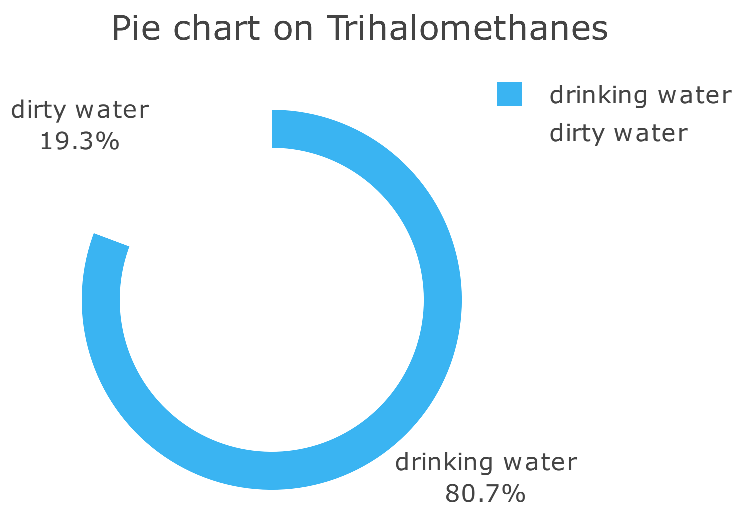
Conductivity:

```
x=c(length(which(data$Conductivity<400)),length(which(data$Conductivity>400)))
txt=c(paste0(round((x[1]/sum(x))*100,2), "%"),paste0(round((x[2]/sum(x))*100,2), "%"))

colfunc <- colorRampPalette(c("#3FFA5E","#3FFA5E", "red","#FA4B3F"))
hist(data$Conductivity,breaks = seq(from=0,to=800,length=50),col=colfunc(50),main = "frequency diagram o
abline(v=400,col="blue")
#locator()
text(90, 120, txt[1])
text(570, 134, txt[2])
text(90, 110, "safe")
text(570, 120, "unsafe")
```



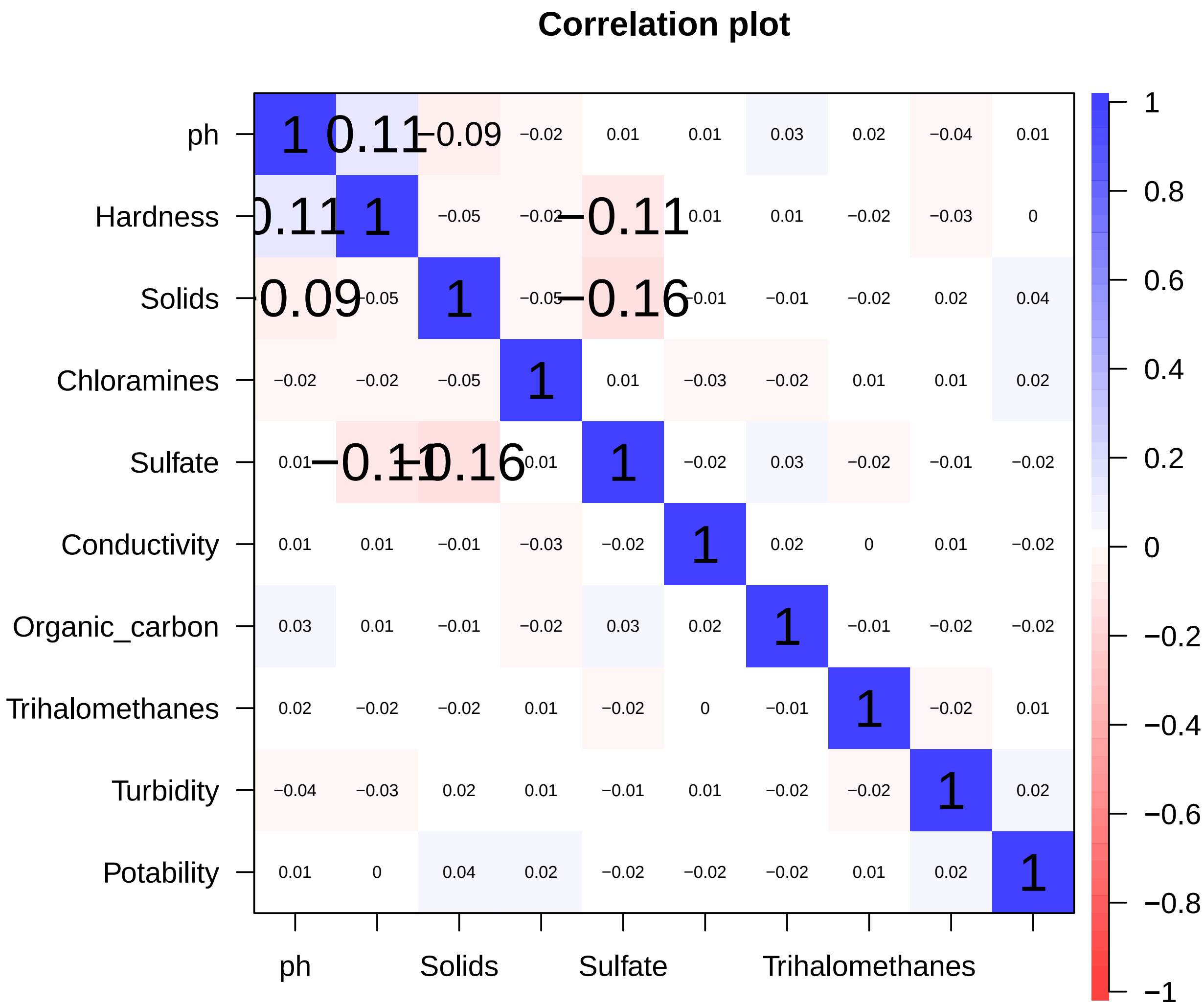
Trihalomethanes:

```
label1=c("drinking water","dirty water")
values1=c(length(which( data$Trihalomethanes<80)),length(which(data$Trihalomethanes>80)))
colors <- c('#3AB4F2','#FFFFFF')
fig <- plot_ly(, labels=label1, values=values1,
               textinfo='label+percent',
               insidetextorientation='radial',marker = list(colors = colors))
fig <- fig %>% add_pie(hole = 0.8)
fig <- fig %>% layout(title = "Pie chart on Trihalomethanes")
fig
```



Correlation plot

```
library(psych)


Attaching package:  'psych'
The following objects are masked from 'package:ggplot2':
      %+%, alpha

corPlot(data, cex = 1.2)
```



this correlation plot is not usefull for us.

Fitting a model:

Splitting dataset into training and test dataset and our spliting ratio 70-30,then extract indepnedent(x) and dependent variables(y)

```
#split dataset
library(caTools)
train = sample.split(data, SplitRatio = .70)
train <- subset( data, train==TRUE )[,1:10]
test <- subset( data, train==FALSE)[,1:10]
trainst=scale(train)
testst=scale(test)
```

```
x=trainst[,1:9]
y=trainst[,10]
```

here we can't perform linear model because our response is in binary format ,so we will use GLM(gerenalized linear model).