607 Project 2 - Water Quality

Mark Schmalfeld

10/1/2021

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
library(curl)
```

```
## Using libcurl 7.64.1 with LibreSSL/2.8.3
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter()     masks stats::filter()
## x dplyr::lag()        masks stats::lag()
## x readr::parse_date() masks curl::parse_date()
```

```
library(stringr)
```

Portable water quality evaluation project

Background from the kagle site. “https:/www.kaggle.com/artimule/drinking-water-probability”

Context 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 drinkingwaterpotability.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 the acidic or alkaline condition of water status. WHO has recommended the 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 an unwanted taste and diluted color in the appearance of water. This is the important parameter for the use of water. The water with a high TDS value indicates that water is highly mineralized. The desirable limit for TDS is 500 mg/l and the maximum limit is 1000 mg/l which is prescribed for drinking purposes. 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. An increase in ions concentration enhances the electrical conductivity of water. Generally, the amount of dissolved solids in water determines 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 exceed 400 microS/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 that 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 are 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 the light-emitting properties of water and the test is used to indicate the quality of waste discharge with respect to the 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. Inspiration Contaminated water and poor sanitation are linked to the transmission of diseases such as cholera, diarrhea, dysentery, hepatitis A, typhoid, and polio. Absent, inadequate, or inappropriately managed water and sanitation services expose individuals to preventable health risks. This is particularly the case in health care facilities where both patients and staff are placed at additional risk of infection and disease when water, sanitation, and hygiene services are lacking.

Load file from github.

```
urlfile<-"https://raw.githubusercontent.com/schmalmr/607_Fall_2021_Project_2_Water/main/Potable_Water_C
water <- read_csv(url(urlfile))
```

```
## Rows: 3276 Columns: 10
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (10): ph, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic_...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
water<-as_tibble(water)
glimpse(water)
```

```
## Rows: 3,276
## Columns: 10
## $ ph              <dbl> NA, 3.716080, 8.099124, 8.316766, 9.092223, 5.584087, ~
## $ Hardness        <dbl> 204.8905, 129.4229, 224.2363, 214.3734, 181.1015, 188.~
## $ Solids          <dbl> 20791.32, 18630.06, 19909.54, 22018.42, 17978.99, 2874~
## $ Chloramines     <dbl> 7.300212, 6.635246, 9.275884, 8.059332, 6.546600, 7.54~
## $ Sulfate         <dbl> 368.5164, NA, NA, 356.8861, 310.1357, 326.6784, 393.66~
## $ Conductivity    <dbl> 564.3087, 592.8854, 418.6062, 363.2665, 398.4108, 280.~
## $ Organic_carbon  <dbl> 10.379783, 15.180013, 16.868637, 18.436525, 11.558279,~
## $ Trihalomethanes <dbl> 86.99097, 56.32908, 66.42009, 100.34167, 31.99799, 54.~
## $ Turbidity       <dbl> 2.963135, 4.500656, 3.055934, 4.628771, 4.075075, 2.55~
## $ Potability      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
```

Based on the specifications for each variables classification as potable water or not potable water - setup a truth table for what passes as potable water and what fails for each specification.

Review using summary:

Interestingly, it does not appear any of the datasets pass as potable water making it less clear the given rating of potable (1) and non-potable(0) in the file. There must be some other factors being used to decide.

For example the solids and organic carbon have 100% or almost 100% failure to meet the potable water standard set in the document. Nothing passes the test from a specification point of view - so this data may represent what is actually being consumed as potable water vs what is currently not consumed due to not being classified as potable. (not clear from the data)

```
water<- water %>%
  mutate(water,between(ph,6.5,8.5))
water<- water %>%
  mutate(water,between(Solids,0,1000))
water<-water %>%
  mutate(water,between(Chloramines,0,4))
water<- water %>%
  mutate(water,between(Sulfate,3,30))
water<- water %>%
  mutate (water,between(Conductivity,0,400))
water<- water %>%
  mutate(water,between(Organic_carbon,0,2))
water<- water %>%
  mutate(water, between(Trihalomethanes,0,80))
water<- water %>%
  mutate(water,between(Turbidity,0,5))

summary(water)
```

```
##        ph            Hardness          Solids         Chloramines
##  Min.   : 0.000   Min.   : 47.43   Min.   :  320.9   Min.   : 0.352
##  1st Qu.: 6.093   1st Qu.:176.85   1st Qu.:15666.7   1st Qu.: 6.127
##  Median : 7.037   Median :196.97   Median :20927.8   Median : 7.130
##  Mean   : 7.081   Mean   :196.37   Mean   :22014.1   Mean   : 7.122
##  3rd Qu.: 8.062   3rd Qu.:216.67   3rd Qu.:27332.8   3rd Qu.: 8.115
```

```
##  Max.   :14.000   Max.   :323.12   Max.   :61227.2   Max.   :13.127
##  NA's   :491
##     Sulfate       Conductivity   Organic_carbon  Trihalomethanes
##  Min.   :129.0   Min.   :181.5   Min.   : 2.20   Min.   :  0.738
##  1st Qu.:307.7   1st Qu.:365.7   1st Qu.:12.07   1st Qu.: 55.845
##  Median :333.1   Median :421.9   Median :14.22   Median : 66.622
##  Mean   :333.8   Mean   :426.2   Mean   :14.28   Mean   : 66.396
##  3rd Qu.:360.0   3rd Qu.:481.8   3rd Qu.:16.56   3rd Qu.: 77.337
##  Max.   :481.0   Max.   :753.3   Max.   :28.30   Max.   :124.000
##  NA's   :781                                     NA's   :162
##    Turbidity       Potability     between(ph, 6.5, 8.5)
##  Min.   :1.450   Min.   :0.0000   Mode :logical
##  1st Qu.:3.440   1st Qu.:0.0000   FALSE:1457
##  Median :3.955   Median :0.0000   TRUE :1328
##  Mean   :3.967   Mean   :0.3901   NA's :491
##  3rd Qu.:4.500   3rd Qu.:1.0000
##  Max.   :6.739   Max.   :1.0000
##
##  between(Solids, 0, 1000) between(Chloramines, 0, 4) between(Sulfate, 3, 30)
##  Mode :logical            Mode :logical              Mode :logical
##  FALSE:3274               FALSE:3187                 FALSE:2495
##  TRUE :2                  TRUE :89                   NA's :781
##
##
##
##
##  between(Conductivity, 0, 400) between(Organic_carbon, 0, 2)
##  Mode :logical                 Mode :logical
##  FALSE:1962                    FALSE:3276
##  TRUE :1314
##
##
##
##
##  between(Trihalomethanes, 0, 80) between(Turbidity, 0, 5)
##  Mode :logical                   Mode :logical
##  FALSE:602                       FALSE:314
##  TRUE :2512                      TRUE :2962
##  NA's :162
##
##
##
```

```
view(water)
```

The potable and non-potable water is seperated for the purposed of creating box plots to evaluate the data. The fundemental differences are non-obvious when we look at the two classifications for the majority of the data. Essentially, both appear to be very similar with some slight differences in the number of outliers or size of the QTL ranges.

The outliers are not consistently higher or more out of spec for the potable or non-potable water. Additionally, the size of the QTR3-QTR1 range is also not consistently tighter for the potable or the non-potable. If anything, potable water seems to have a larger QTR3-QTR1 range more often in the data set.

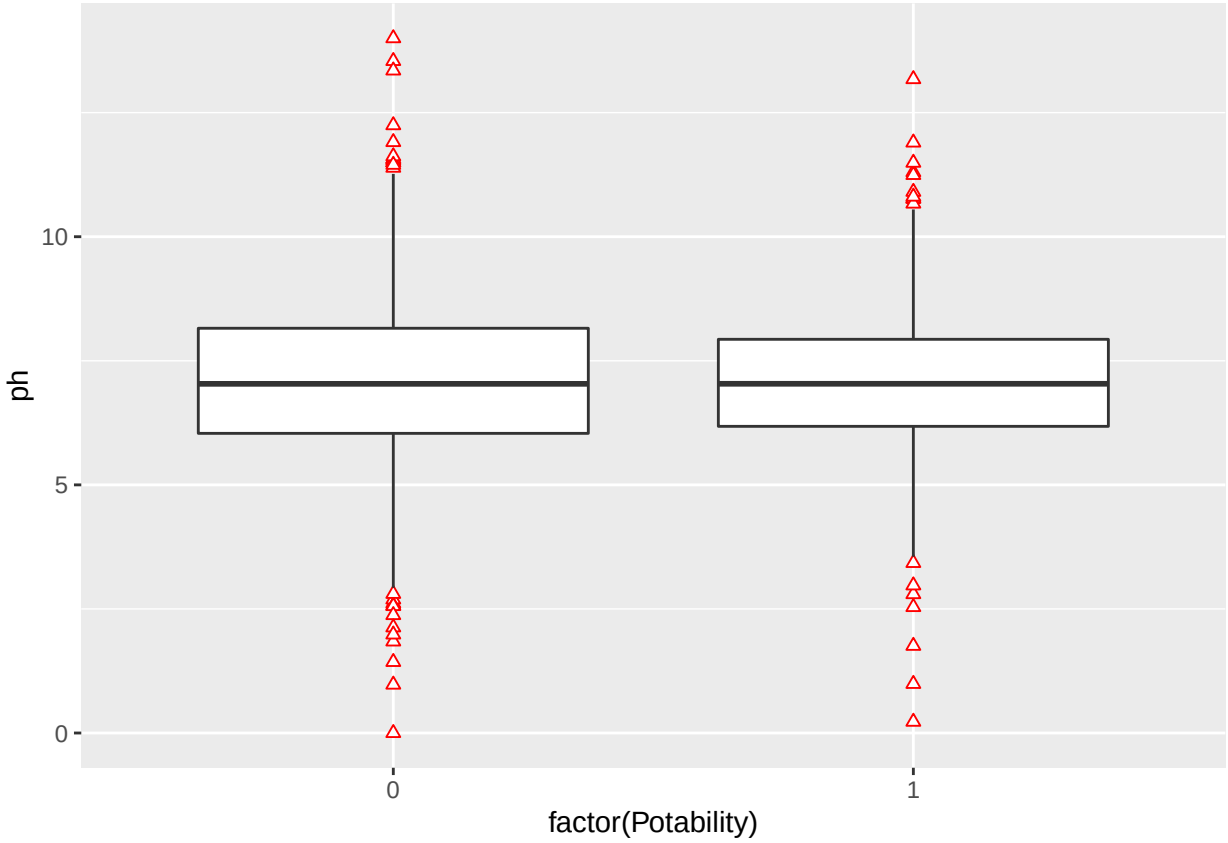
The source of the data would be interesting as well as more information on why the criteria of potable or non-

potable was assigned. Possibilities: It represents actually use or non-use of water sources, it is intentially not true with a target to get a system to properly assess the water, or maybe the water has additional treatment downstream for those that fail to meet potable water criteria but are labeled potable while those that are non-potable can not be treated any further.

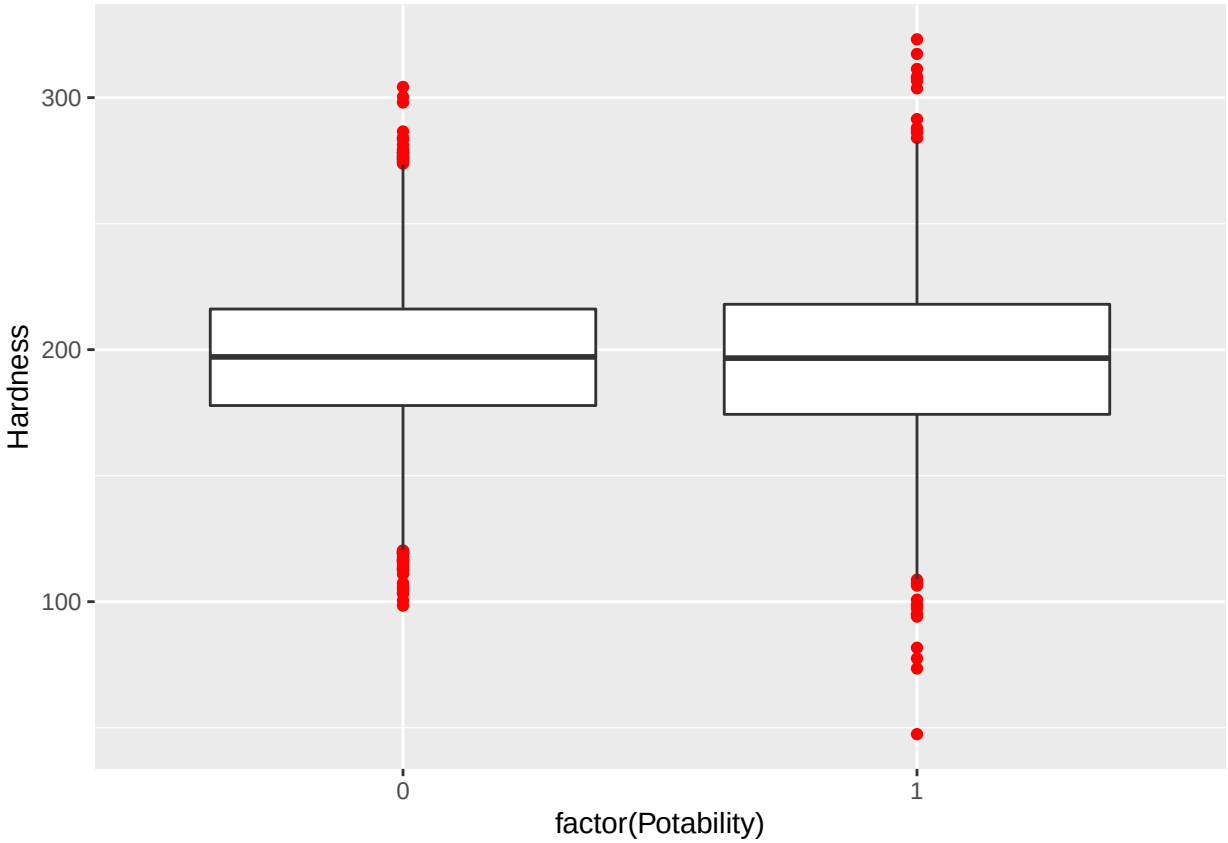
The histograms following by potable and non-potable criteria also do not reveal anything more enlighting to me.

```
ggplot(water,aes(x=factor(Potability),y=ph))+geom_boxplot(outlier.colour = "red",outlier.shape = 24)
```

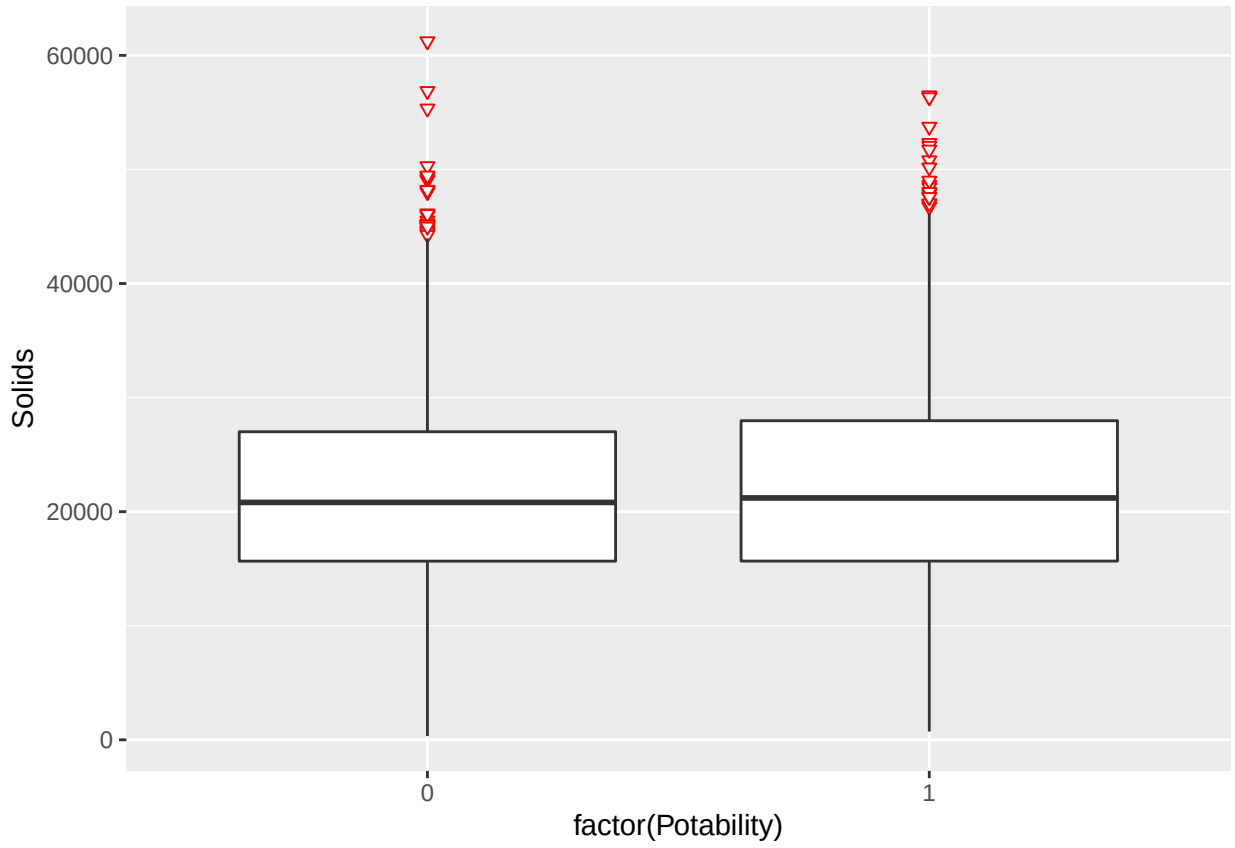
```
## Warning: Removed 491 rows containing non-finite values (stat_boxplot).
```



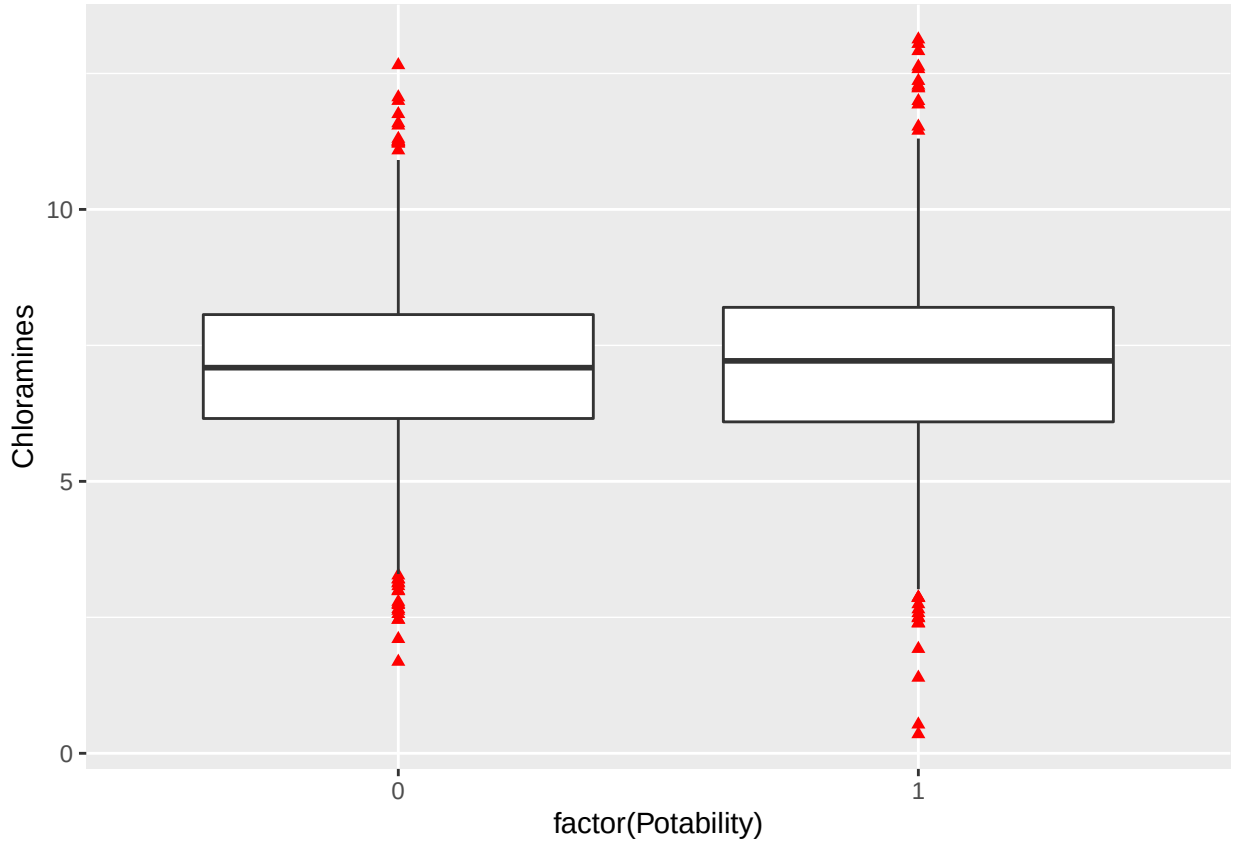

```
ggplot(water,aes(x=factor(Potability),y=Hardness))+geom_boxplot(outlier.colour = "red",outlier.shape =
```




```
ggplot(water,aes(x=factor(Potability),y=Solids))+geom_boxplot(outlier.colour = "red",outlier.shape = 25
```

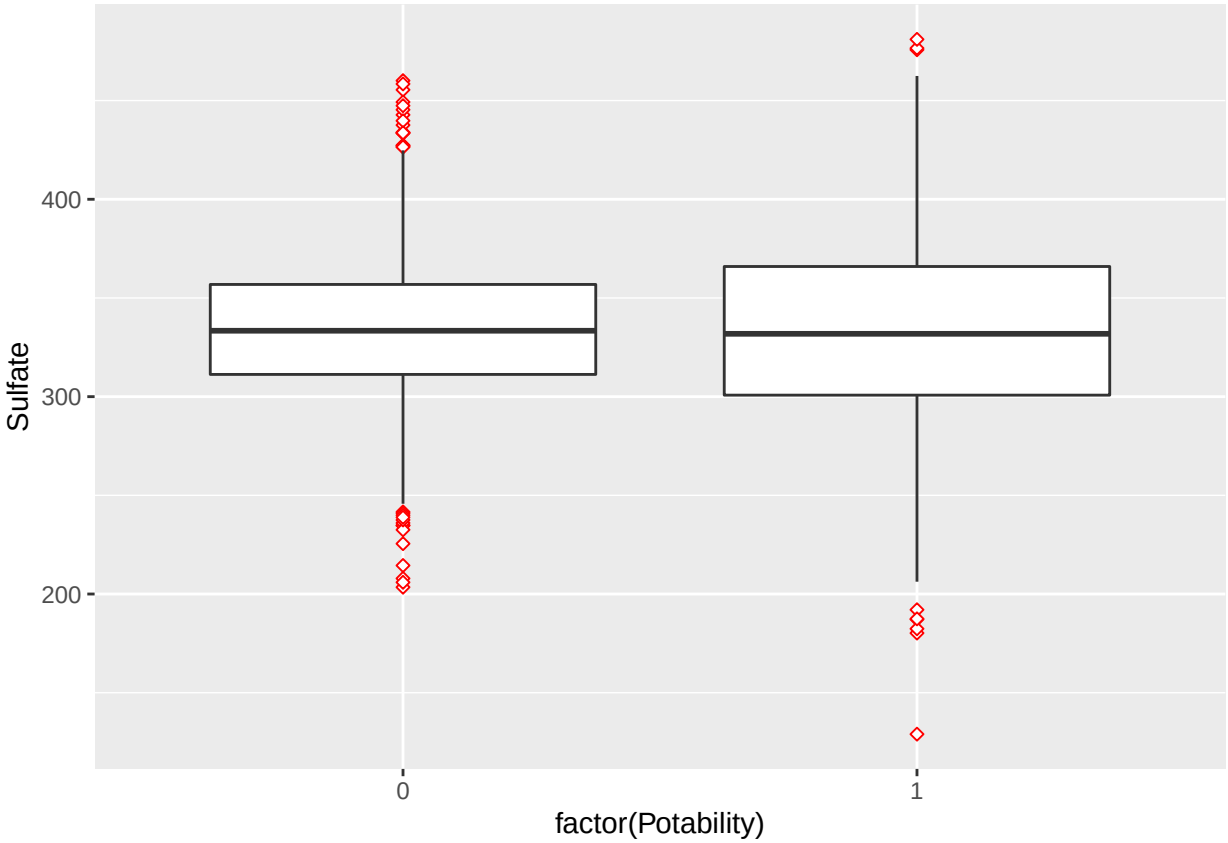



```
ggplot(water,aes(x=factor(Potability),y=Chloramines))+geom_boxplot(outlier.colour = "red",outlier.shape
```

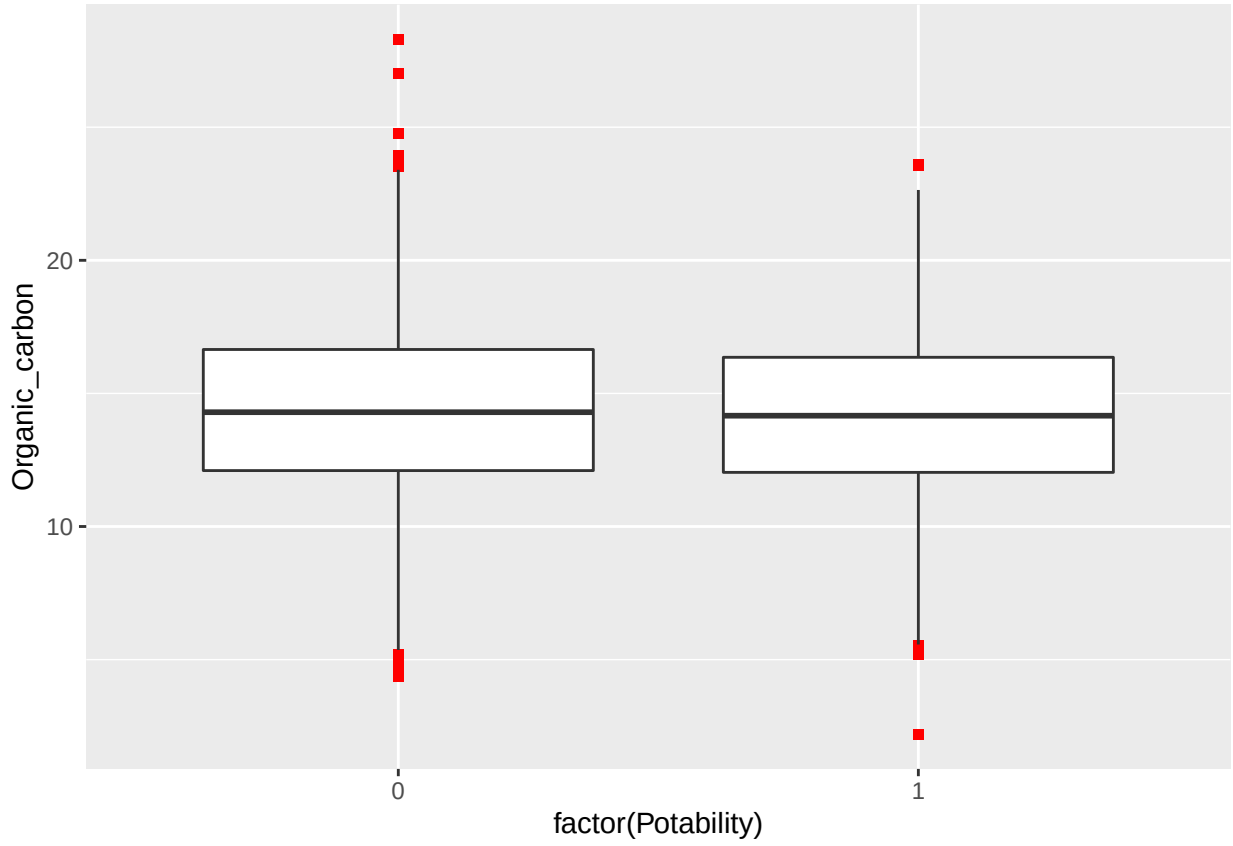



```
ggplot(water,aes(x=factor(Potability),y=Sulfate))+geom_boxplot(outlier.colour = "red",outlier.shape = 2
```

```
## Warning: Removed 781 rows containing non-finite values (stat_boxplot).
```

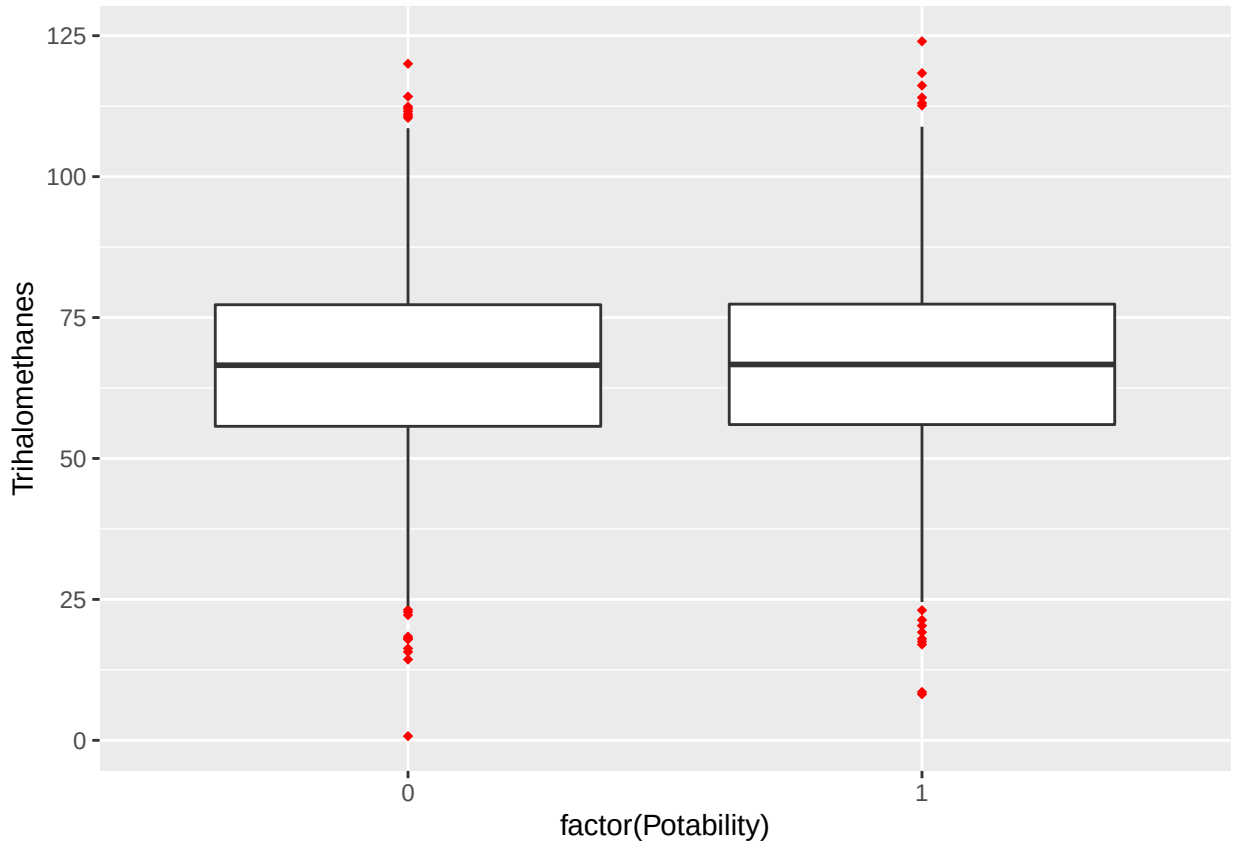



```
ggplot(water,aes(x=factor(Potability),y=Organic_carbon))+geom_boxplot(outlier.colour = "red",outlier.sha
```

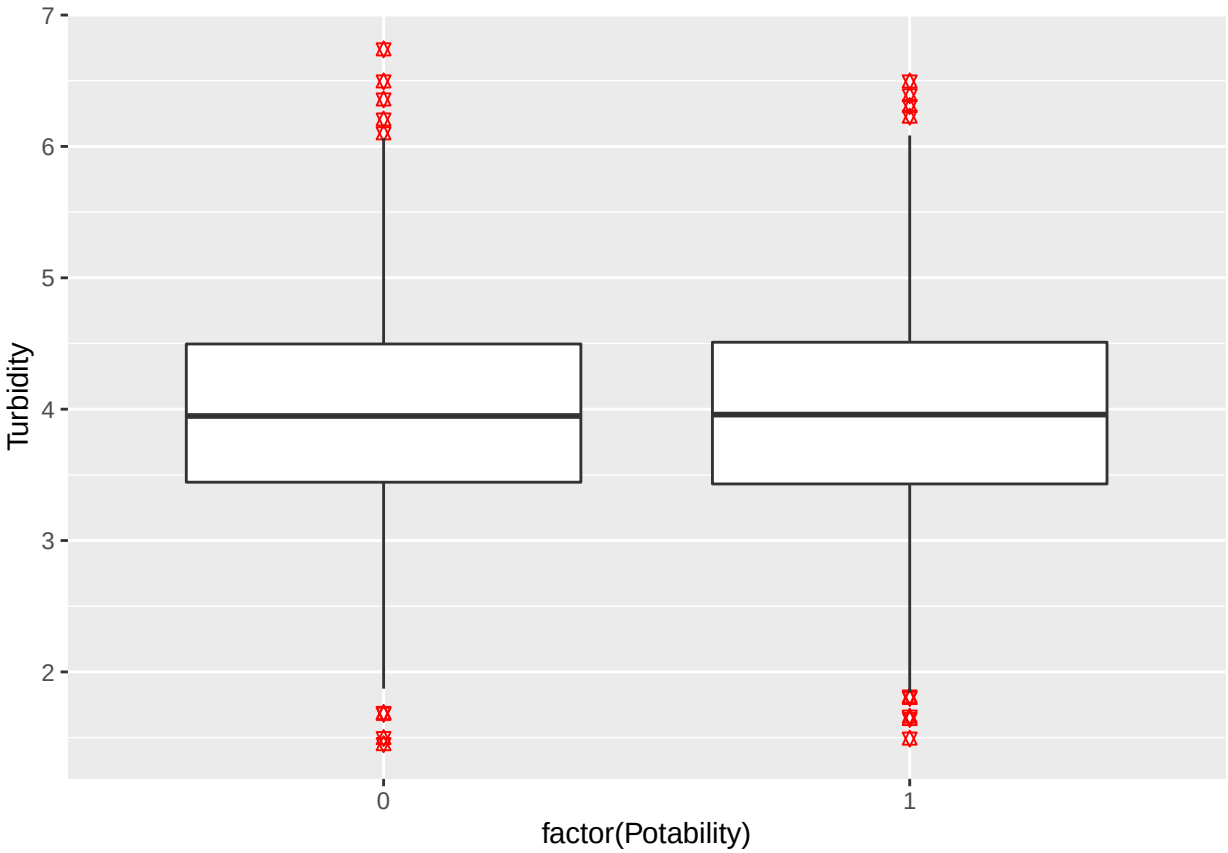



```
ggplot(water,aes(x=factor(Potability),y=Trihalomethanes))+geom_boxplot(outlier.colour = "red",outlier.sh
```

```
## Warning: Removed 162 rows containing non-finite values (stat_boxplot).
```




```
ggplot(water,aes(x=factor(Potability),y=Turbidity))+geom_boxplot(outlier.colour = "red",outlier.shape =
```



This section continued to analyze the data. Below the data is summarized with a comparison of the statistics for the file classified as Potable and non-Potable water. The fundemental gap is essentially all of the water fails to meet the required specifications for potable water. The main parameter that seems to support the potable water is the Turbidity measurement but even this has some out of specifications in the potable category.

SUMMARY of the Potable Water Results (using input data for potable criteria) ph Hardness Solids Chloramines
Min. : 0.2275 Min. : 47.43 Min. : 728.8 Min. : 0.352
1st Qu.: 6.1793 1st Qu.:174.33 1st Qu.:15669.0 1st Qu.: 6.094
Median : 7.0367 Median :196.63 Median :21199.4 Median : 7.215
Mean : 7.0738 Mean :195.80 Mean :22384.0 Mean : 7.169
3rd Qu.: 7.9331 3rd Qu.:218.00 3rd Qu.:27973.2 3rd Qu.: 8.199
Max. :13.1754 Max. :323.12 Max. :56488.7 Max. :13.127
NA's :177
Sulfate Conductivity Organic_carbon Trihalomethanes
Min. :129.0 Min. :201.6 Min. : 2.20 Min. : 8.176
1st Qu.:300.8 1st Qu.:360.9 1st Qu.:12.03 1st Qu.: 56.014
Median :331.8 Median :420.7 Median :14.16 Median : 66.678
Mean :332.6 Mean :425.4 Mean :14.16 Mean : 66.540
3rd Qu.:365.9 3rd Qu.:484.2 3rd Qu.:16.36 3rd Qu.: 77.381
Max. :481.0 Max. :695.4 Max. :23.60 Max. :124.000
NA's :293 NA's :55
Turbidity Potability Min. :1.492 Min. :1
1st Qu.:3.431 1st Qu.:1
Median :3.959 Median :1
Mean :3.968 Mean :1

3rd Qu.:4.510 3rd Qu.:1
Max. :6.494 Max. :1

Specifications attainment for the POTABLE water as classified in the file

between(ph, 6.5, 8.5) between(Solids, 0, 1000) Mode :logical Mode :logical
FALSE:518 FALSE:1277
TRUE :583 TRUE :1
NA's :177

between(Chloramines, 0, 4) between(Sulfate, 3, 30) between(Conductivity, 0, 400) Mode :logical Mode :logical Mode :logical
FALSE:1242 FALSE:985 FALSE:762
TRUE :36 NA's :293 TRUE :516

between(Organic_carbon, 0, 2) between(Trihalomethanes, 0, 80) Mode :logical Mode :logical
FALSE:1278 FALSE:238
TRUE :985
NA's :55

between(Turbidity, 0, 5) Mode :logical
FALSE:117
TRUE :1161

SUMMARY of the NON-Potable water (based on the input criteria ranking) ph Hardness Solids Chloramines
Min. : 0.000 Min. : 98.45 Min. : 320.9 Min. : 1.684
1st Qu.: 6.038 1st Qu.:177.82 1st Qu.:15663.1 1st Qu.: 6.156
Median : 7.035 Median :197.12 Median :20809.6 Median : 7.090
Mean : 7.085 Mean :196.73 Mean :21777.5 Mean : 7.092
3rd Qu.: 8.156 3rd Qu.:216.12 3rd Qu.:27006.2 3rd Qu.: 8.066
Max. :14.000 Max. :304.24 Max. :61227.2 Max. :12.653
NA's :314
Sulfate Conductivity Organic_carbon Trihalomethanes
Min. :203.4 Min. :181.5 Min. : 4.372 Min. : 0.738
1st Qu.:311.3 1st Qu.:368.5 1st Qu.:12.101 1st Qu.: 55.707
Median :333.4 Median :422.2 Median :14.294 Median : 66.542
Mean :334.6 Mean :426.7 Mean :14.364 Mean : 66.304
3rd Qu.:356.9 3rd Qu.:480.7 3rd Qu.:16.649 3rd Qu.: 77.278
Max. :460.1 Max. :753.3 Max. :28.300 Max. :120.030
NA's :488 NA's :107
Turbidity Potability Min. :1.450 Min. :0
1st Qu.:3.444 1st Qu.:0
Median :3.948 Median :0
Mean :3.966 Mean :0
3rd Qu.:4.496 3rd Qu.:0
Max. :6.739 Max. :0

Specification performance for each category for the NON-Potable water as classified in the file

between(ph, 6.5, 8.5) between(Solids, 0, 1000) Mode :logical Mode :logical
FALSE:939 FALSE:1997
TRUE :745 TRUE :1
NA's :314

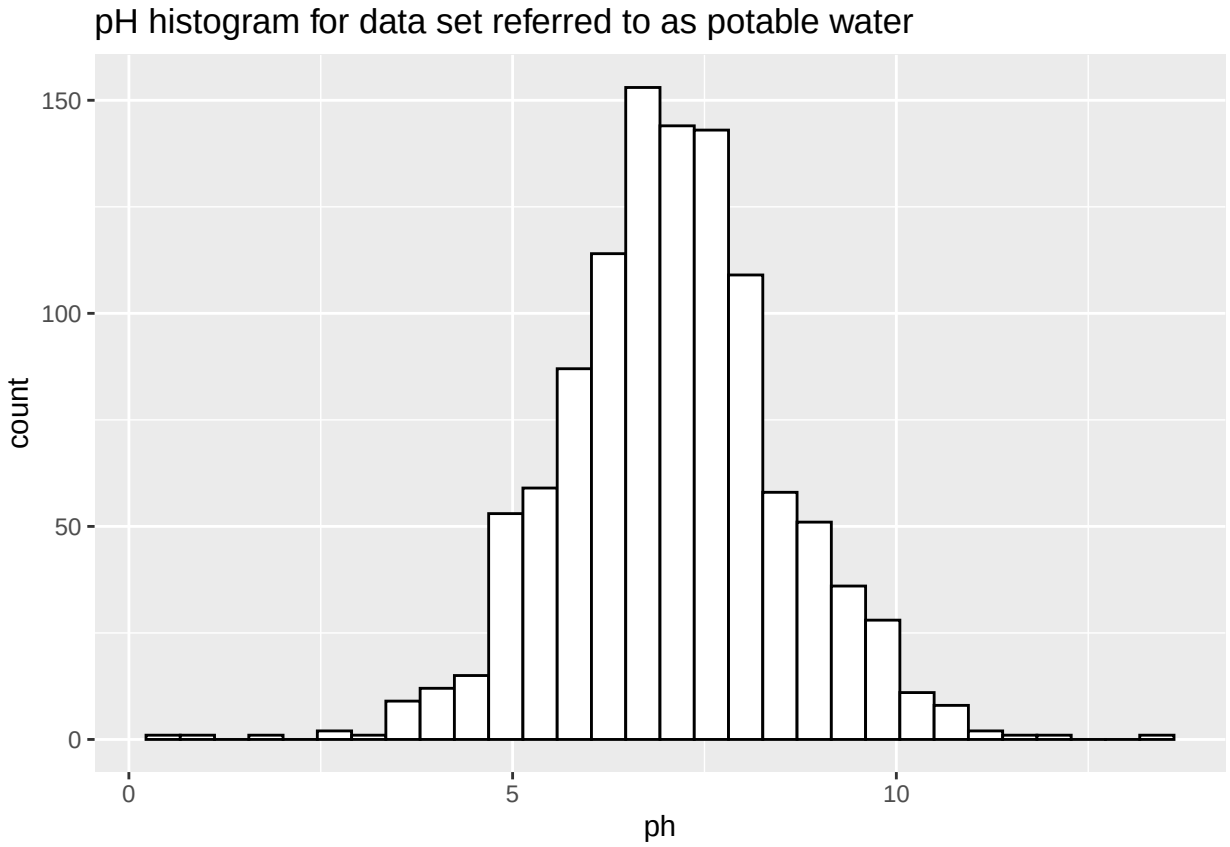
between(Chloramines, 0, 4) between(Sulfate, 3, 30) between(Conductivity, 0, 400) Mode :logical Mode :logical Mode :logical
FALSE:1945 FALSE:1510 FALSE:1200
TRUE :53 NA's :488 TRUE :798

between(Organic_carbon, 0, 2) between(Trihalomethanes, 0, 80) Mode :logical Mode :logical
FALSE:1998 FALSE:364
TRUE :1527
NA's :107

between(Turbidity, 0, 5) Mode :logical
FALSE:197
TRUE :1801

```
waterpotable<-filter(water,Potability==1)
waternotpotable<-filter(water,Potability==0)

summary_waterpotable<-(summary(waterpotable))   #Summary of the potable water based on the input data r
view(data.frame(summary_waterpotable))
summary_waternonpotable<-(summary(waternotpotable)) #Summary of the non-potable water based on the inpu
view(data.frame(summary_waternonpotable))

ggplot(waterpotable,aes(x=ph))+geom_histogram(fill="white", color="black")+labs(title="pH histogram for
```

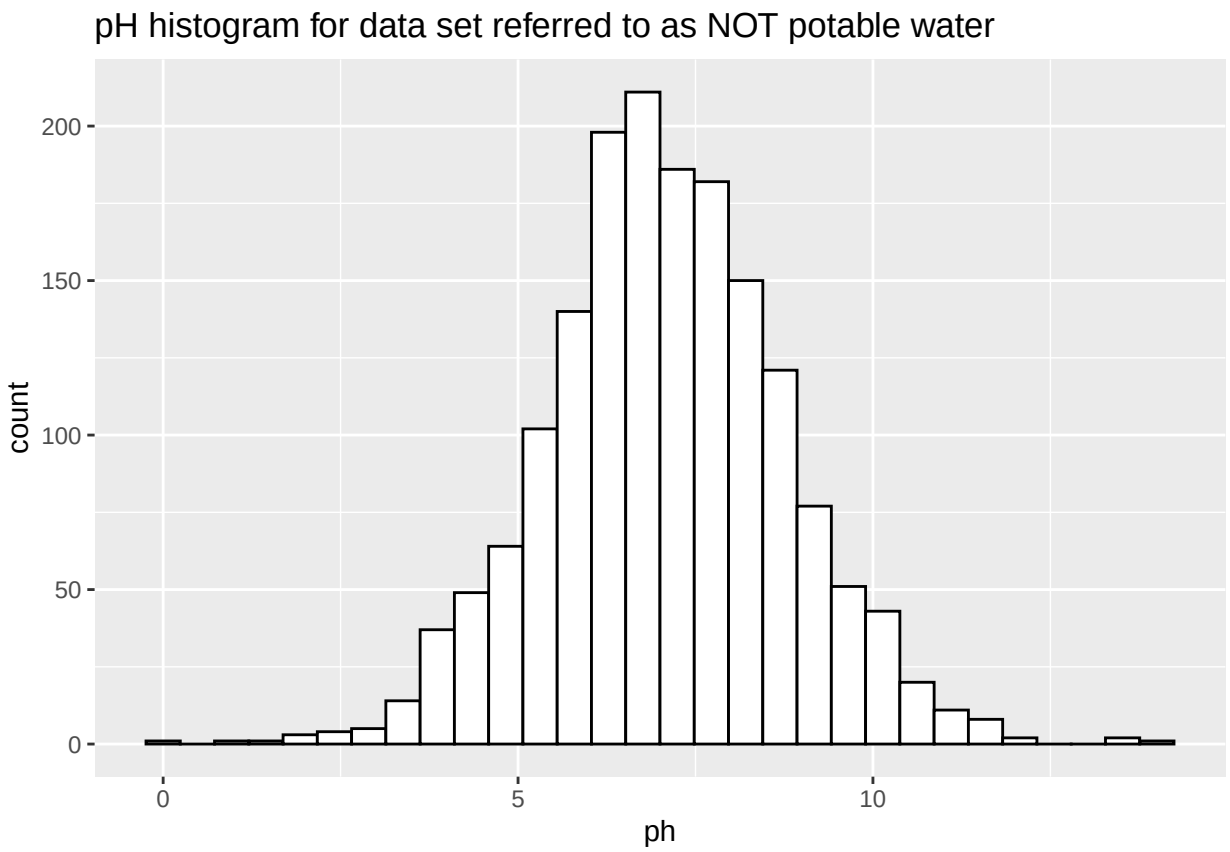
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 177 rows containing non-finite values (stat_bin).
```



```
ggplot(waternotpotable,aes(x=ph))+geom_histogram(fill="white", color="black")+labs(title="pH histogram
```

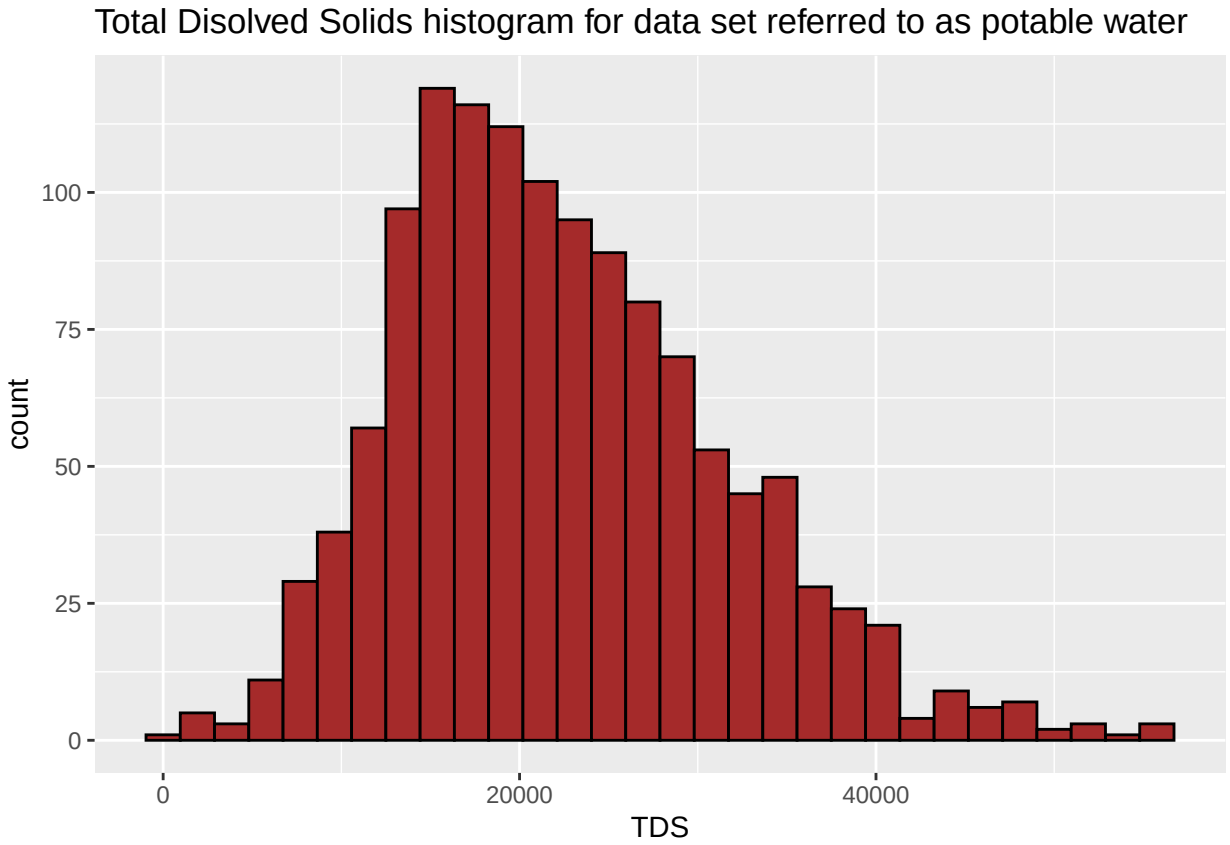
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 314 rows containing non-finite values (stat_bin).
```



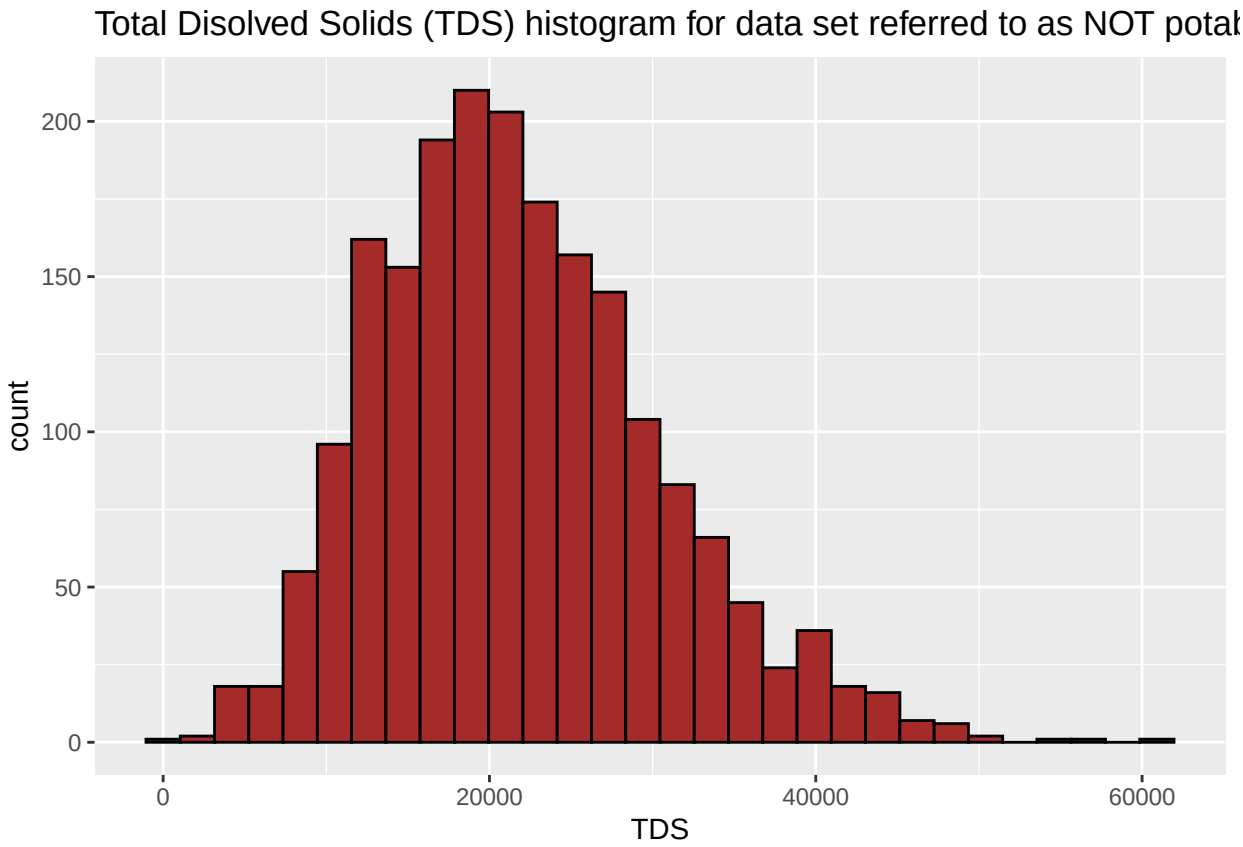

```
ggplot(waterpotable,aes(x=Solids))+geom_histogram(fill="brown", color="black")+labs(title="Total Disolv
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



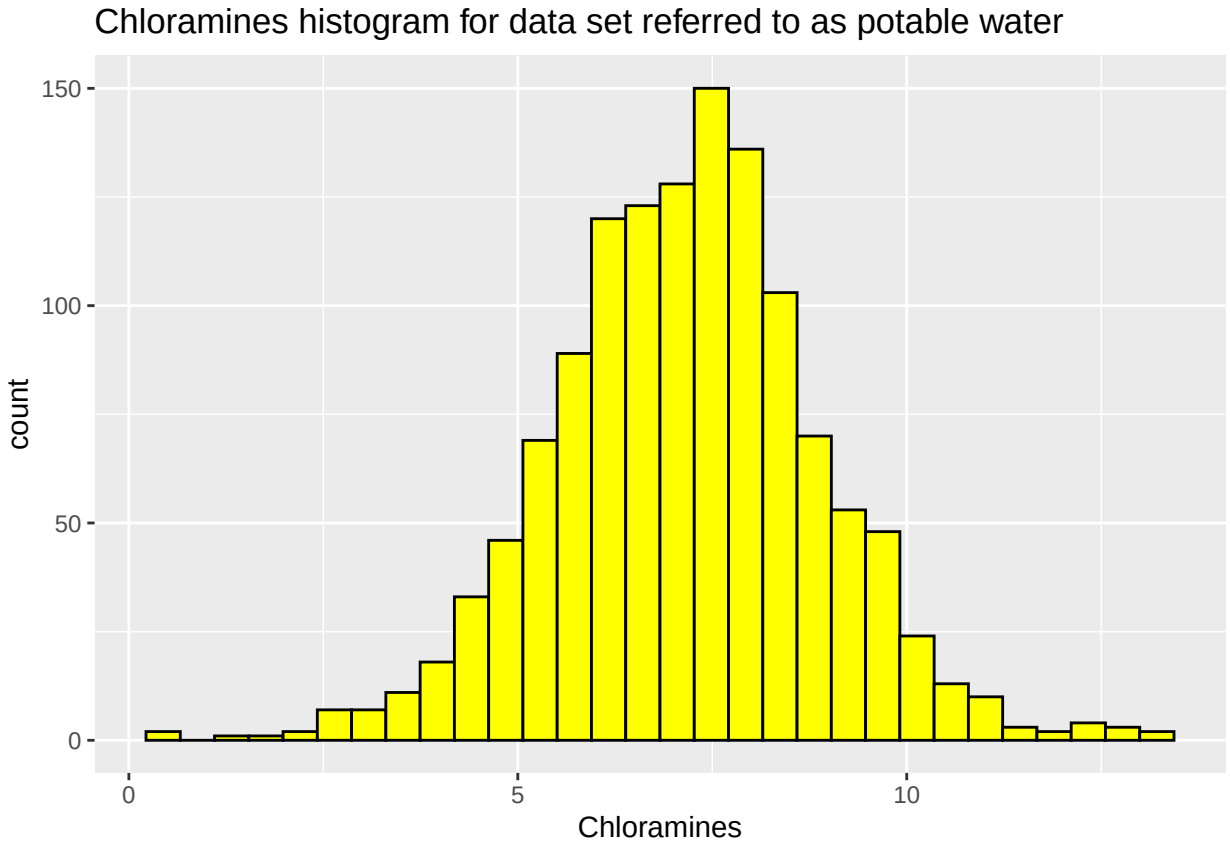

```
ggplot(waternotpotable,aes(x=Solids))+geom_histogram(fill="brown", color="black")+labs(title="Total Dis
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



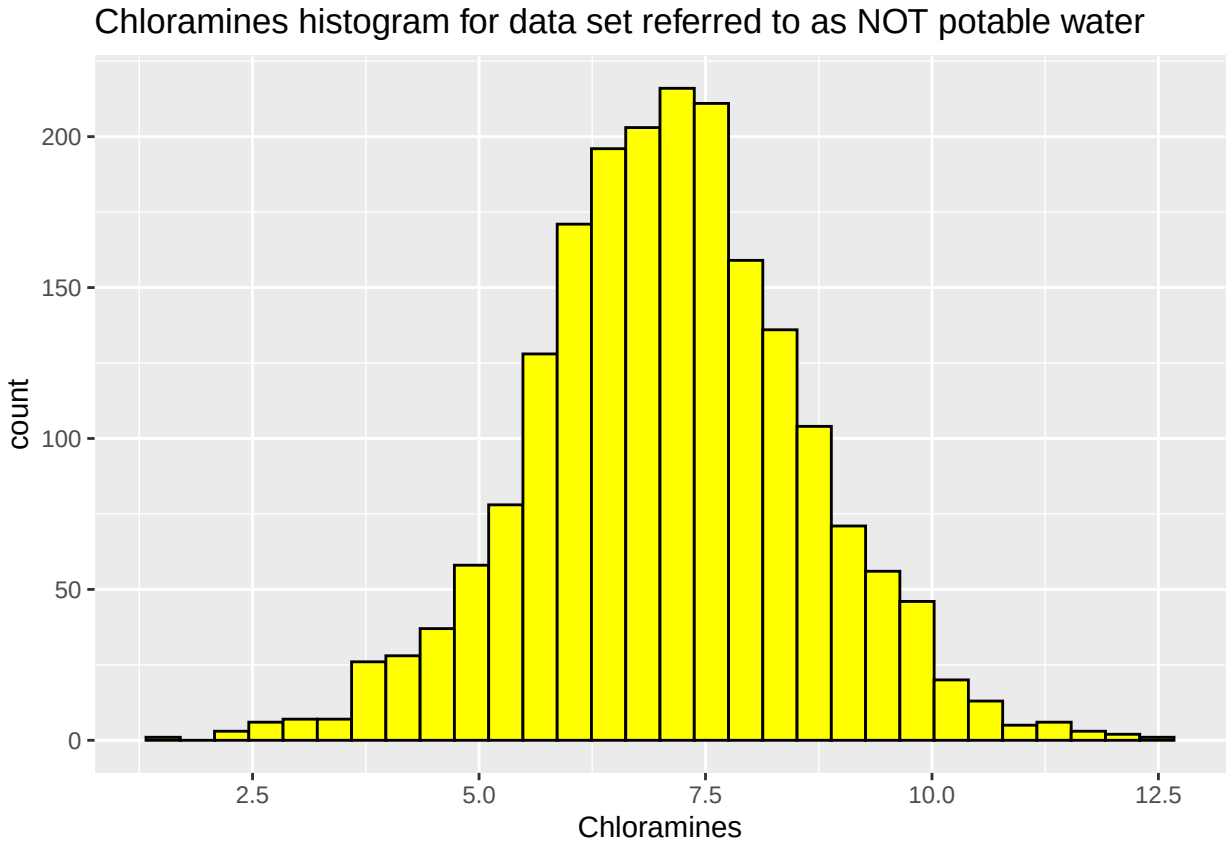

```
ggplot(waterpotable,aes(x=Chloramines))+geom_histogram(fill="yellow", color="black")+labs(title="Chloram
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(waternotpotable,aes(x=Chloramines))+geom_histogram(fill="yellow", color="black")+labs(title="Chl
```

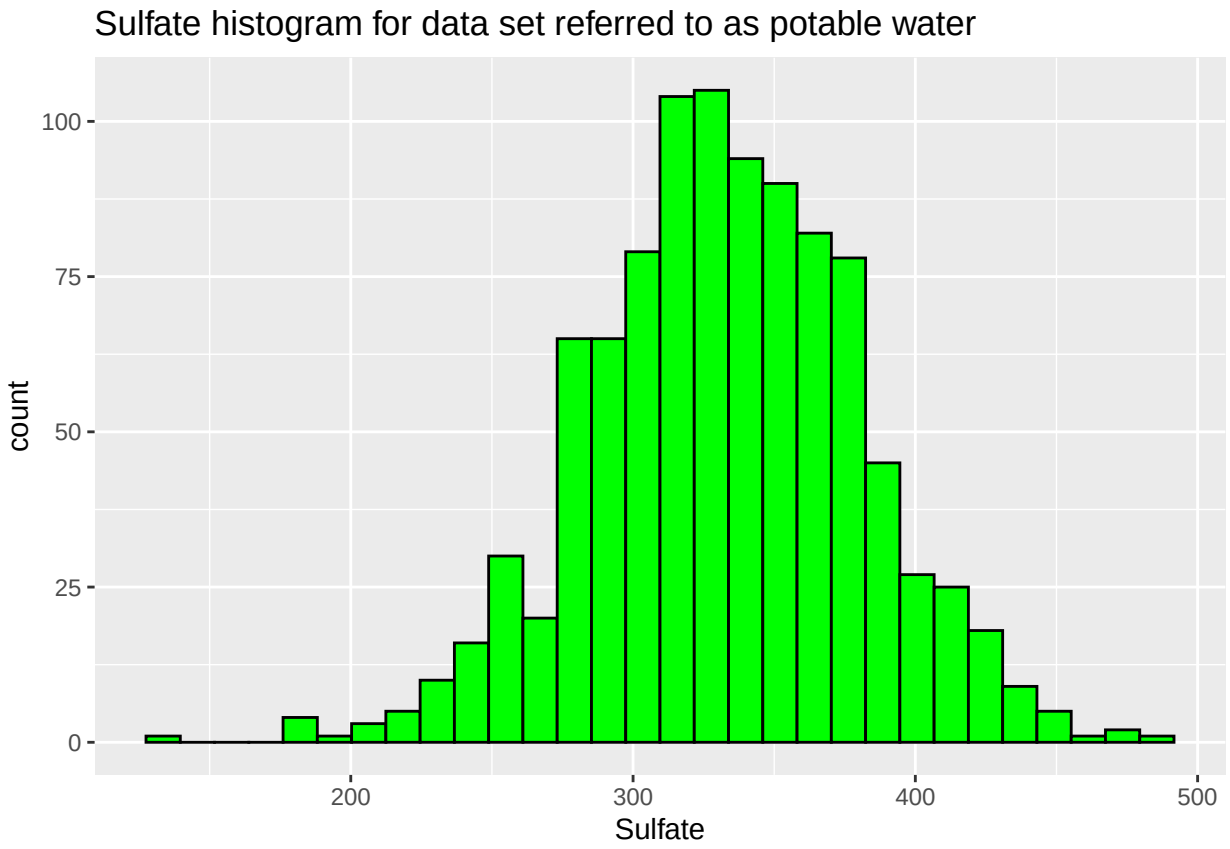
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(waterpotable,aes(x=Sulfate))+geom_histogram(fill="green", color="black")+labs(title="Sulfate his
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

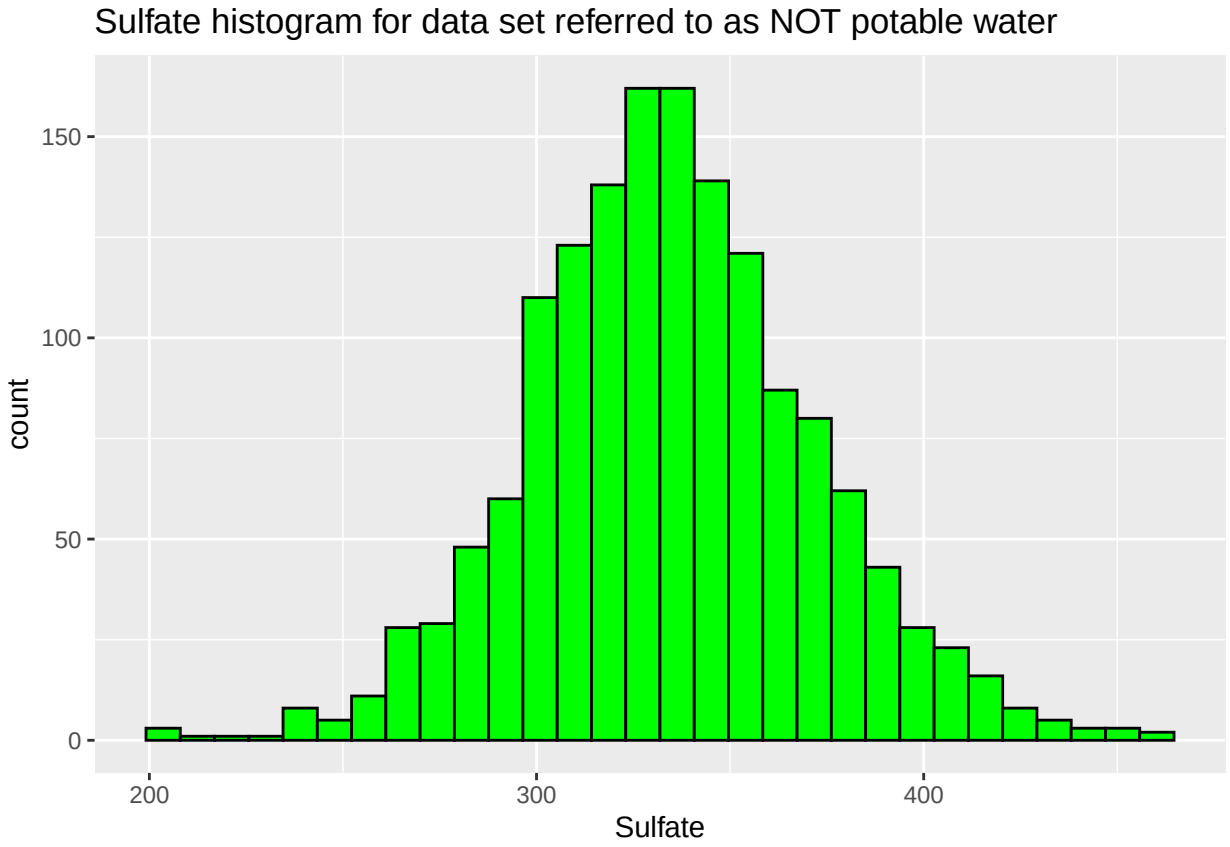
```
## Warning: Removed 293 rows containing non-finite values (stat_bin).
```




```
ggplot(waternotpotable,aes(x=Sulfate))+geom_histogram(fill="green", color="black")+labs(title="Sulfate h
```

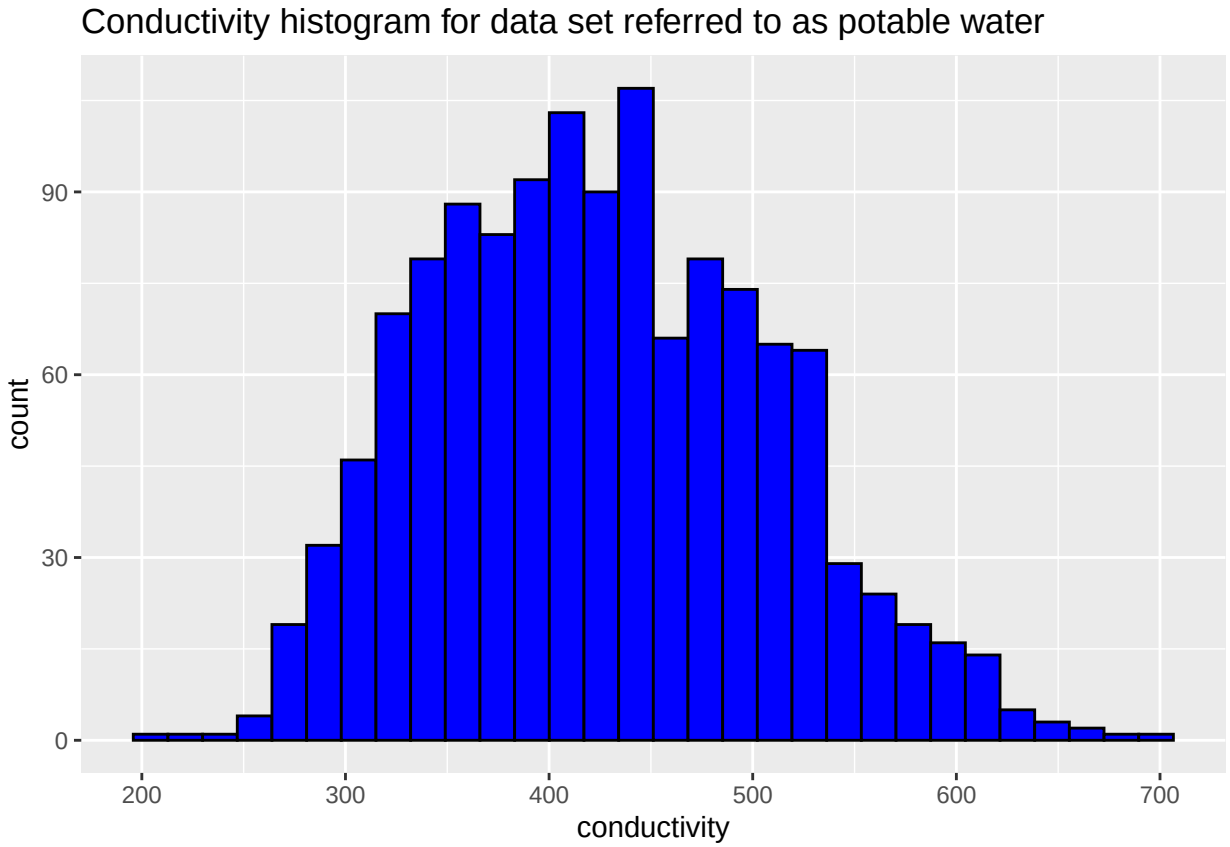
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 488 rows containing non-finite values (stat_bin).
```



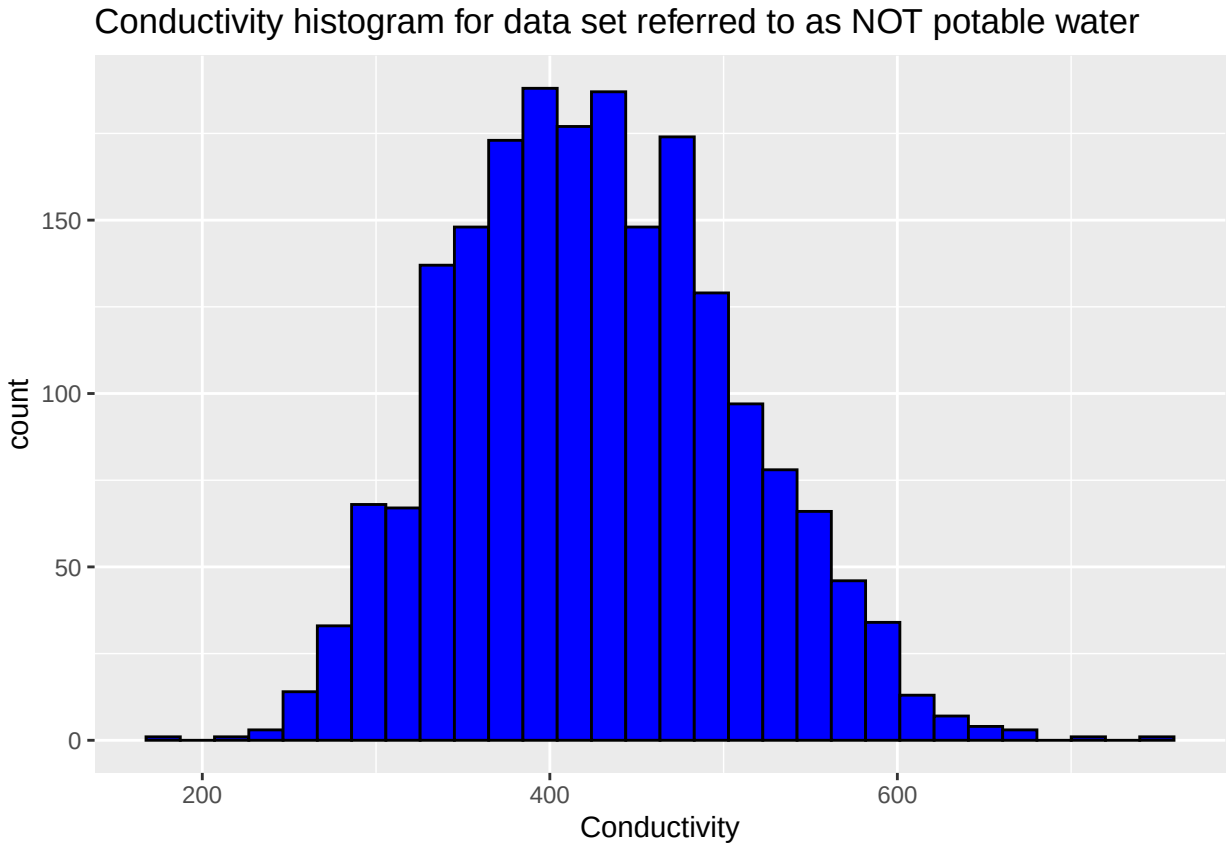

```
ggplot(waterpotable,aes(x=Conductivity))+geom_histogram(fill="blue", color="black")+ labs(title="Conduc
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



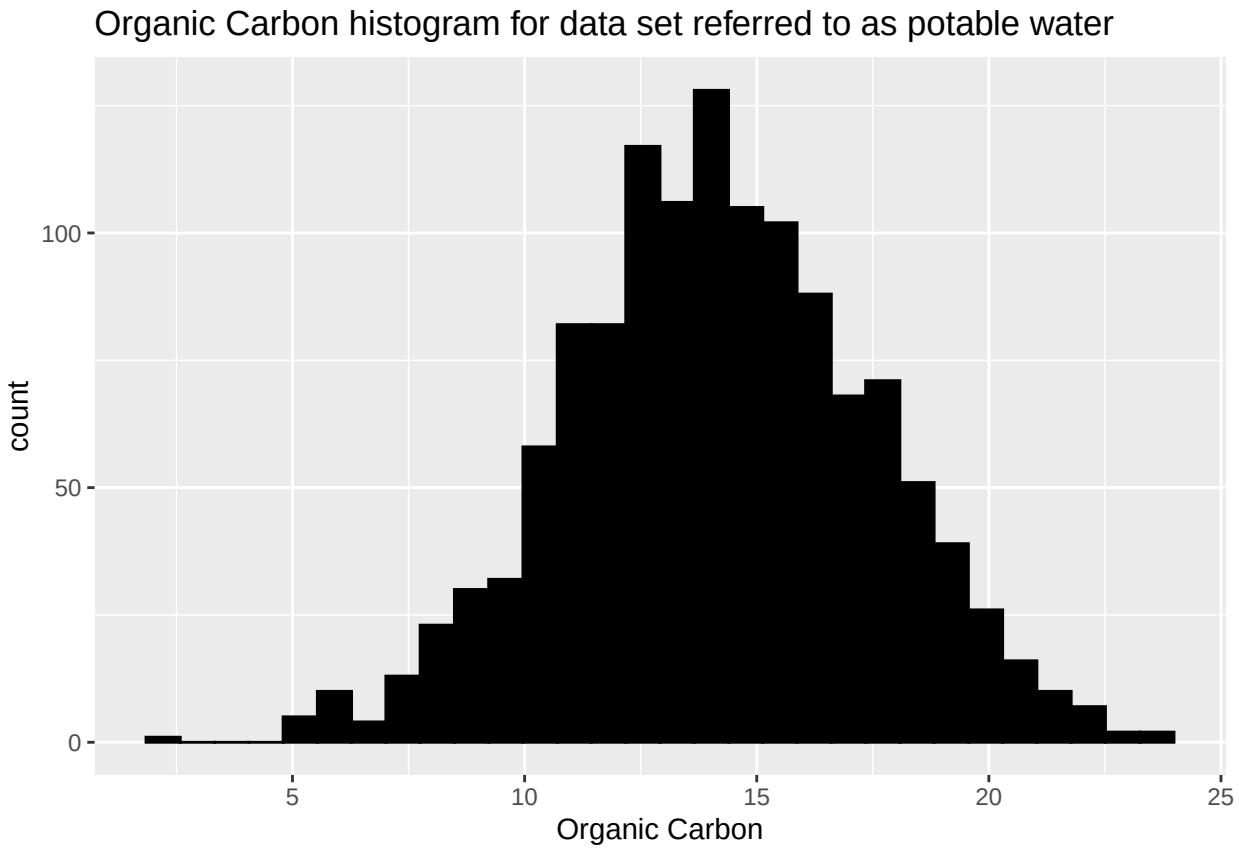

```
ggplot(waternotpotable,aes(x=Conductivity))+geom_histogram(fill="blue", color="black")+labs(title="Condu
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



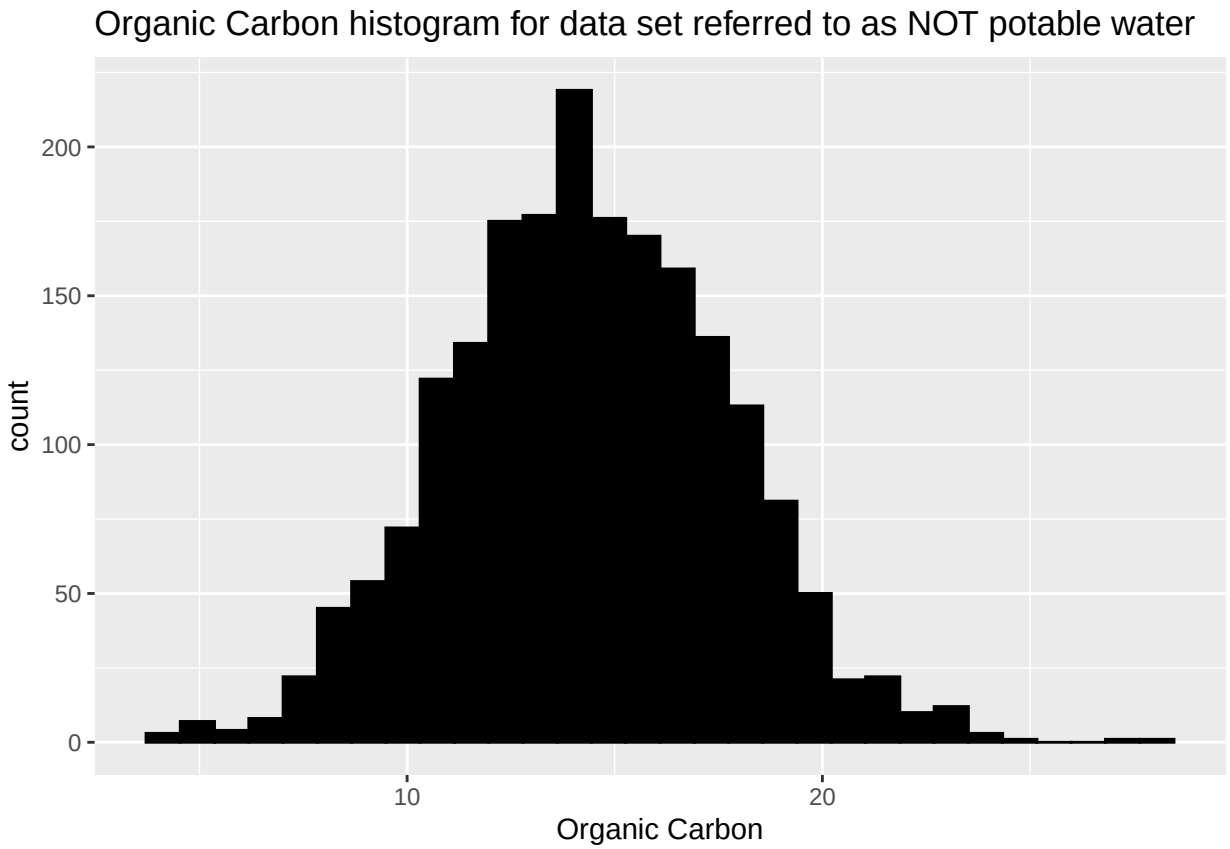

```
ggplot(waterpotable,aes(x=Organic_carbon))+geom_histogram(fill="black", color="black")+labs(title="Orga
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(waternotpotable,aes(x=Organic_carbon))+geom_histogram(fill="black", color="black")+labs(title="O
```

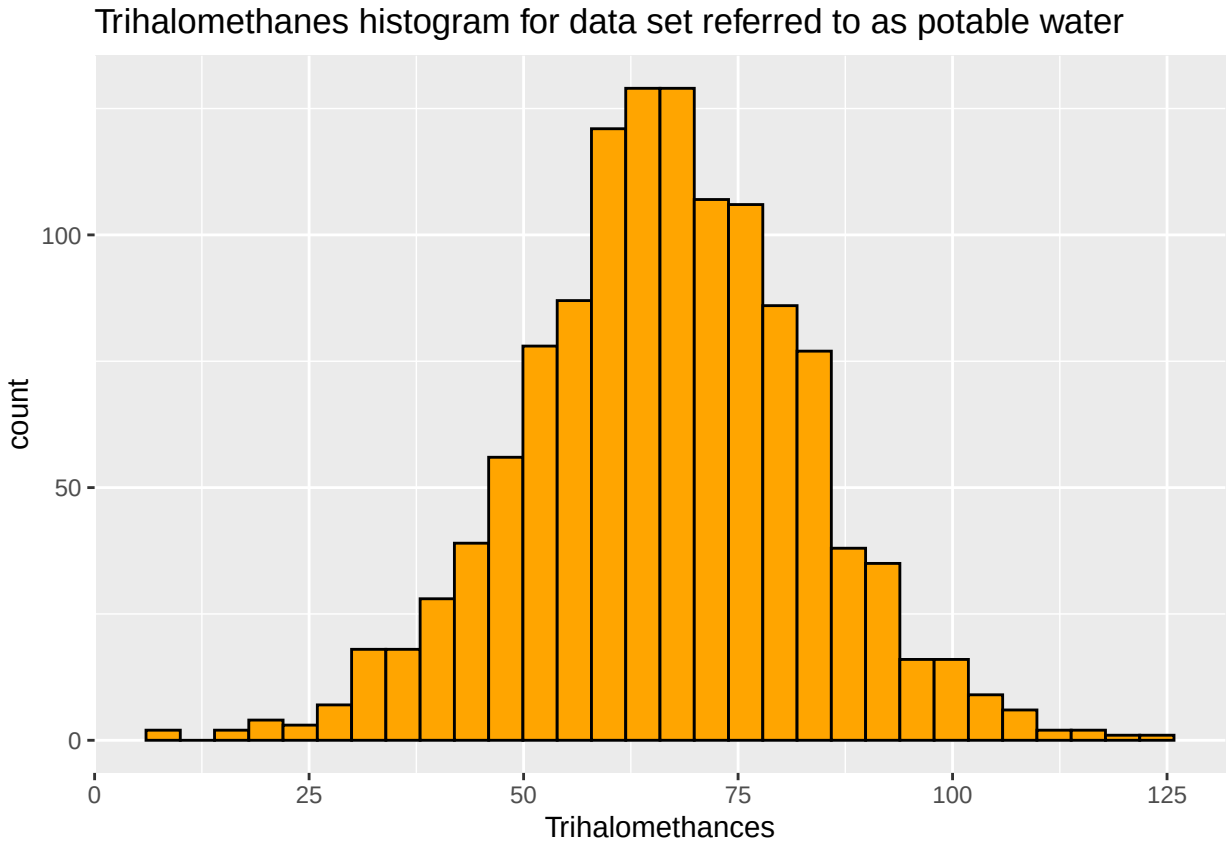
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(waterpotable,aes(x=Trihalomethanes))+geom_histogram(fill="orange", color="black")+labs(title="Tri
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

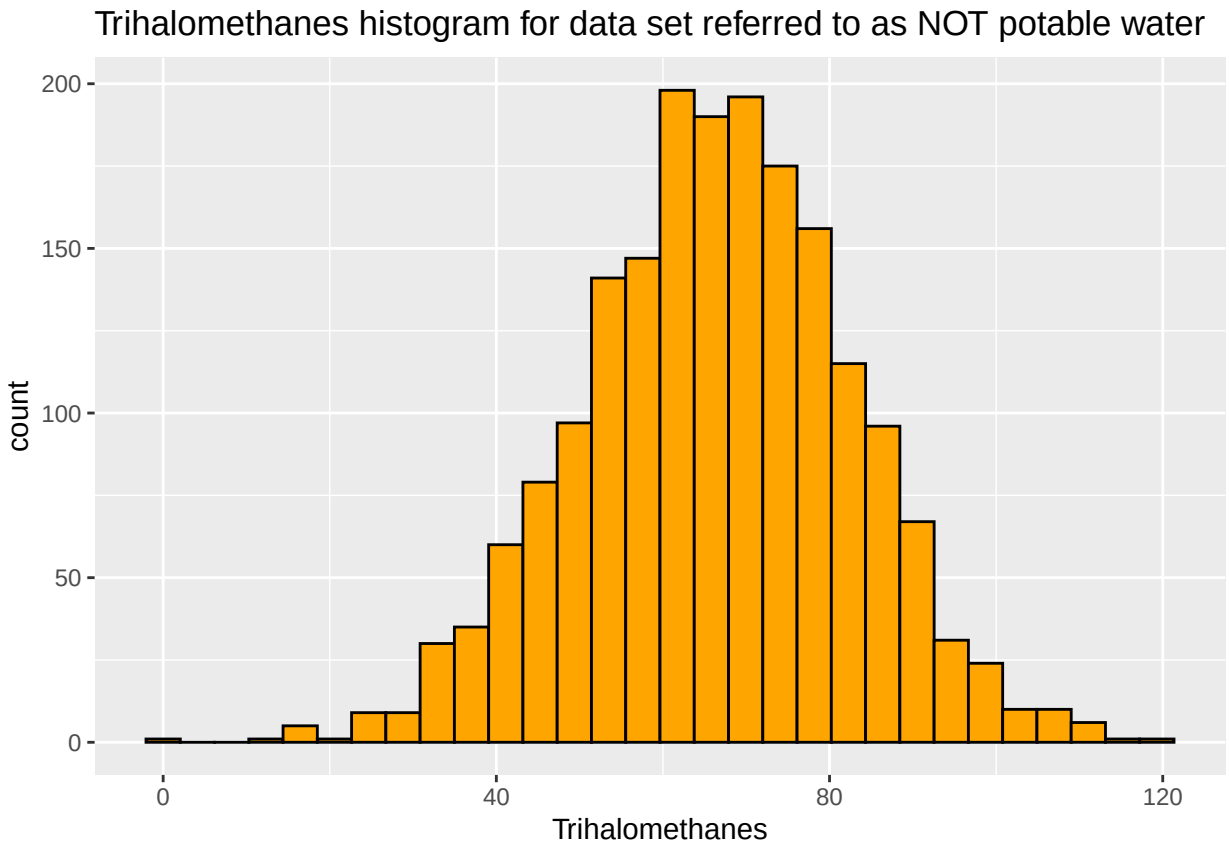
```
## Warning: Removed 55 rows containing non-finite values (stat_bin).
```




```
ggplot(waternotpotable,aes(x=Trihalomethanes))+geom_histogram(fill="orange", color="black")+labs(title=
```

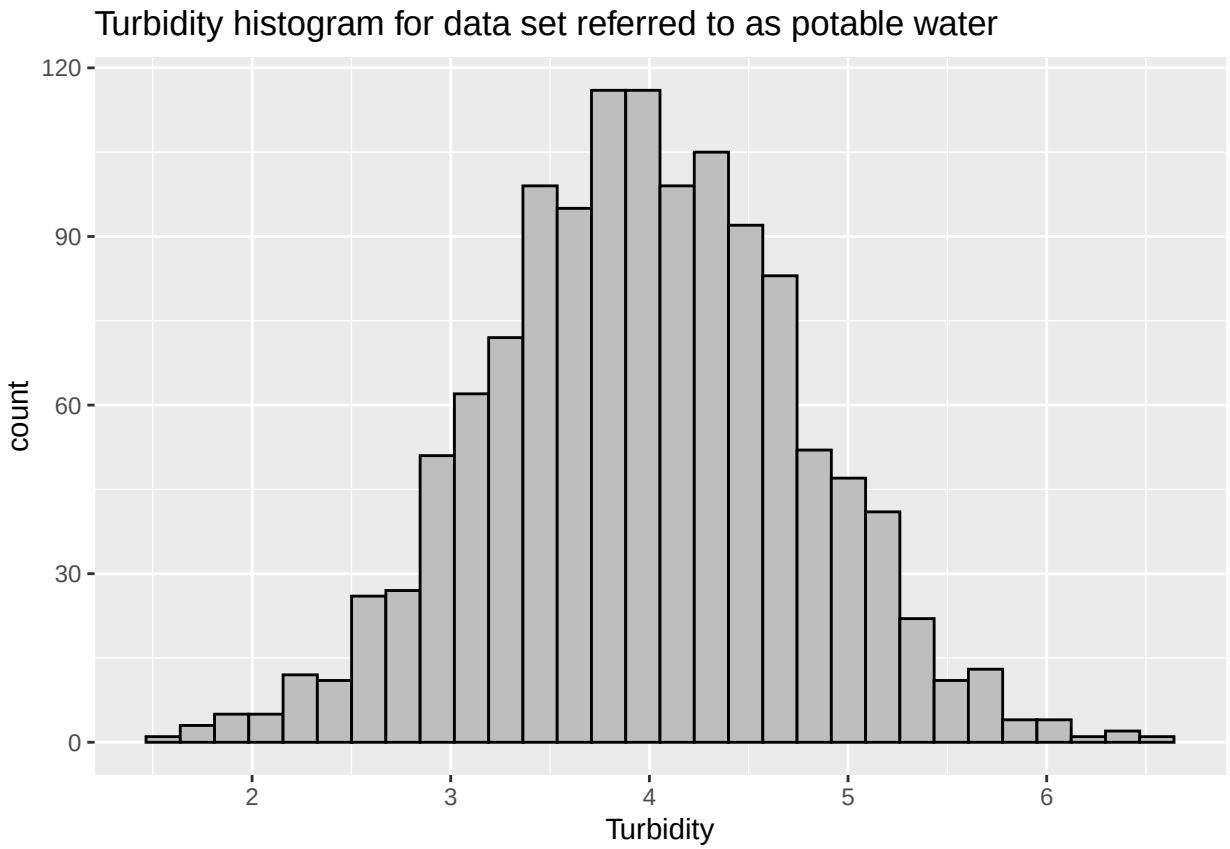
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 107 rows containing non-finite values (stat_bin).
```



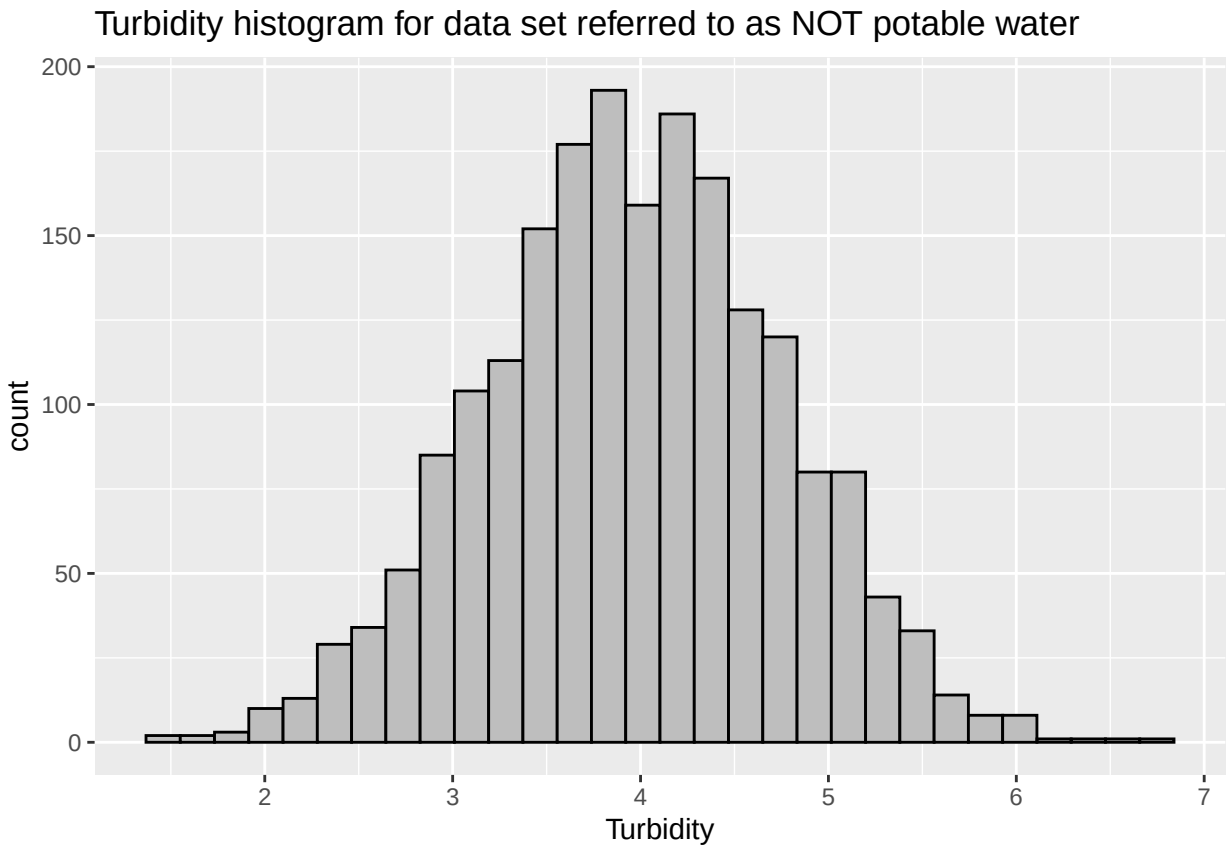

```
ggplot(waterpotable,aes(x=Turbidity))+geom_histogram(fill="grey", color="black")+labs(title="Turbidity h
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



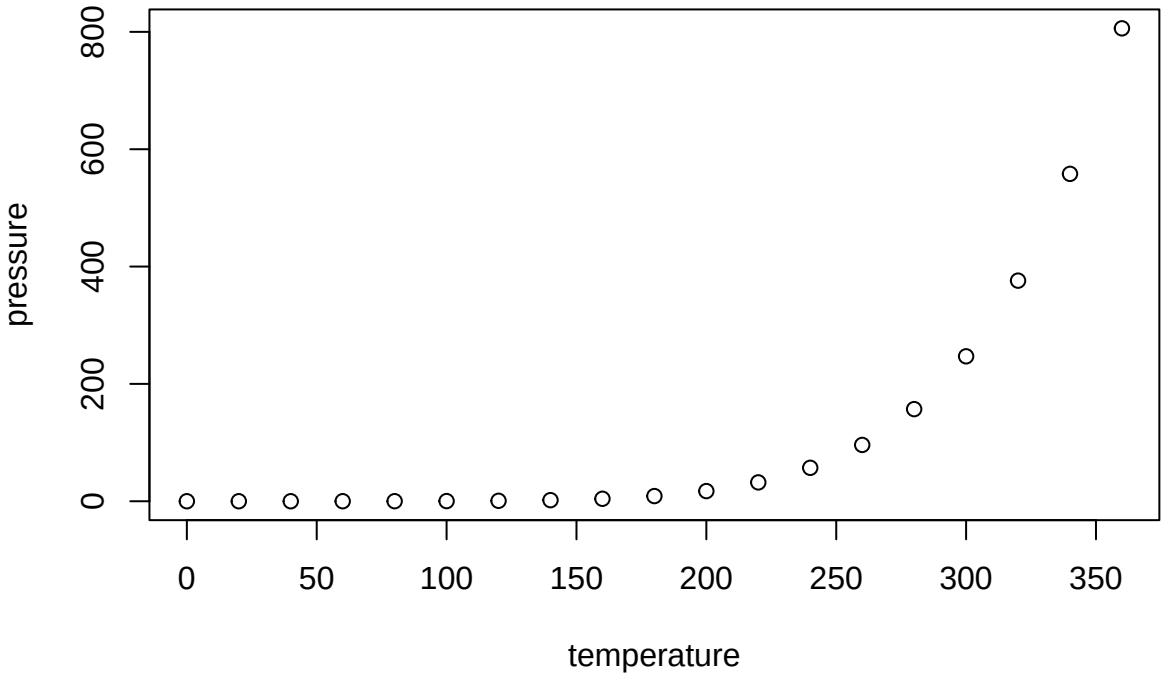

```
ggplot(waternotpotable,aes(x=Turbidity))+geom_histogram(fill="grey", color="black")+labs(title="Turbidi
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.