



Water Quality Assessment

Anoop S Hari
2023-11-26

About the dataset

 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 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 6.52–8.83 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 or 2.7 g/L) and 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 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 exceed 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/L 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 and the amount of chlorine required to treat the water, and the temperature of the water that is being treated. THM levels up to 80 ppb 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.

```
# thead::install.packages(c("tidyverse", "dplyr", "readr", "tidyr"), repos = "https://cran.rstudio.com/")

## package 'tidyverse' successfully unpacked and MD5 sums checked
## package 'dplyr' successfully unpacked and MD5 sums checked
## package 'readr' successfully unpacked and MD5 sums checked
## package 'tidyr' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
## C:\Users\HP\AppData\Local\Temp\RtmpEtE9J87\downloaded_packages

library(tidyverse)
library(dplyr)
library(readr)
library(tidyr)

# Import the dataset
water <- read.csv("F:\\R practice\\Dataset\\water_potability.csv")

# checking the structure and summary of the dataset
str(water)
```

```
## 'data.frame':      3276 obs. of  10 variables:
## $ ph             : num  NA 3.72 8.1 8.32 9.09 ...
## $ Hardness       : num 285 129 224 214 181 ...
## $ Solids         : num 20791 18630 19910 22010 17979 ...
## $ Chloramines    : num  7.3 6.64 9.28 8.06 6.55 ...
## $ Sulfate        : num 369 NA NA 357 310 ...
## $ Conductivity   : num 564 593 419 383 398 ...
## $ Organic_carbon : num 10.4 15.2 16.9 18.4 11.6 ...
## $ Trihalomethanes: num 67 66.3 66.4 100.3 32 ...
## $ Turbidity      : num  2.96 4.5 3.06 4.63 4.06 ...
## $ Potability     : int  0 0 0 0 0 0 0 0 ...
```

```
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 : 7.139
## 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
##
##           Sulfate           Conductivity           Organic_carbon           Trihalomethanes
##  Min.   :129.0   Min.   :181.5   Min.   : 2.20   Min.   : 0.738
## 1st Qu.:307.7   1st Qu.:350.7   1st Qu.:12.07   1st Qu.: 56.846
## 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.:386.0   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           Turbidity           Potability
##  Min.   :1.450   Min.   :0.0000
## 1st Qu.:3.440   1st Qu.:0.0000
## Median :3.955   Median :0.0000
## Mean   :3.967   Mean   :0.3901
## 3rd Qu.:4.500   3rd Qu.:1.0000
## Max.   :6.739   Max.   :1.0000
```

Data Cleaning

```
# Checking for missing values if any in the dataset
colSums(is.na(water))

##           ph           Hardness           Solids           Chloramines           Sulfate
##           491              0              0              0              781
## Conductivity Organic_carbon Trihalomethanes Turbidity Potability
##           0              0              162              0              0

# It seems there are some missing values in columns 'ph', 'Sulfate', 'Trihalomethanes'.

# Missing values were imputed with the mean.
water$Sulfate[is.na(water$Sulfate)] <- mean(water$Sulfate, na.rm = TRUE)

water$ph[is.na(water$ph)] <- mean(water$ph, na.rm = TRUE)

water$Trihalomethanes[is.na(water$Trihalomethanes)] <- mean(water$Trihalomethanes, na.rm = TRUE)

# Confirming the replacement of missing values with mean values.
any(is.na(water))

## [1] FALSE
```

```
summary(water)
```

```
##           ph           Hardness           Solids           Chloramines
##  Min.   : 0.000   Min.   : 47.43   Min.   : 320.9   Min.   : 0.352
## 1st Qu.: 6.278   1st Qu.:176.85   1st Qu.:15666.7   1st Qu.: 6.127
## Median : 7.081   Median :196.97   Median :20927.8   Median : 7.139
## Mean   : 7.081   Mean   :196.37   Mean :22014.1   Mean   : 7.122
## 3rd Qu.: 7.878   3rd Qu.:216.67   3rd Qu.:27332.8   3rd Qu.: 8.115
## Max.   :14.000   Max.   :323.12   Max.   :61227.2   Max.   :13.127
##
##           Sulfate           Conductivity           Organic_carbon           Trihalomethanes
##  Min.   :129.0   Min.   :181.5   Min.   : 2.20   Min.   : 0.738
## 1st Qu.:337.1   1st Qu.:385.7   1st Qu.:12.07   1st Qu.: 56.846
## Median :333.8   Median :421.9   Median :14.22   Median : 66.396
## Mean   :333.8   Mean   :426.2   Mean   :14.28   Mean   : 66.396
## 3rd Qu.:356.4   3rd Qu.:481.8   3rd Qu.:16.56   3rd Qu.: 76.667
## Max.   :481.0   Max.   :753.3   Max.   :28.30   Max.   :124.000
##
##           Turbidity           Potability
##  Min.   :1.450   Min.   :0.0000
## 1st Qu.:3.440   1st Qu.:0.0000
## Median :3.955   Median :0.0000
## Mean   :3.967   Mean   :0.3901
## 3rd Qu.:4.500   3rd Qu.:1.0000
## Max.   :6.739   Max.   :1.0000
```

Analysis

```
# Replacing the values 0 & 1 in Potability column to "Potable" and "Not Potable".
water.filtered <- water %>%
  mutate(Potability = ifelse(Potability == 1, "Potable", "Not Potable"))

# Finding out the percentage of potable and not potable water.
perc.pot <- water.filtered %>%
  group_by(Potability) %>%
  summarise(Count = n()) %>%
  mutate(Perc = (Count / sum(Count)) * 100)

print(perc.pot)

## # A tibble: 2 × 3
##   Potability Count Perc
##   <chr>      <int> <dbl>
## 1 Not Potable 1998  61.0
## 2 Potable    1278  39.0
```

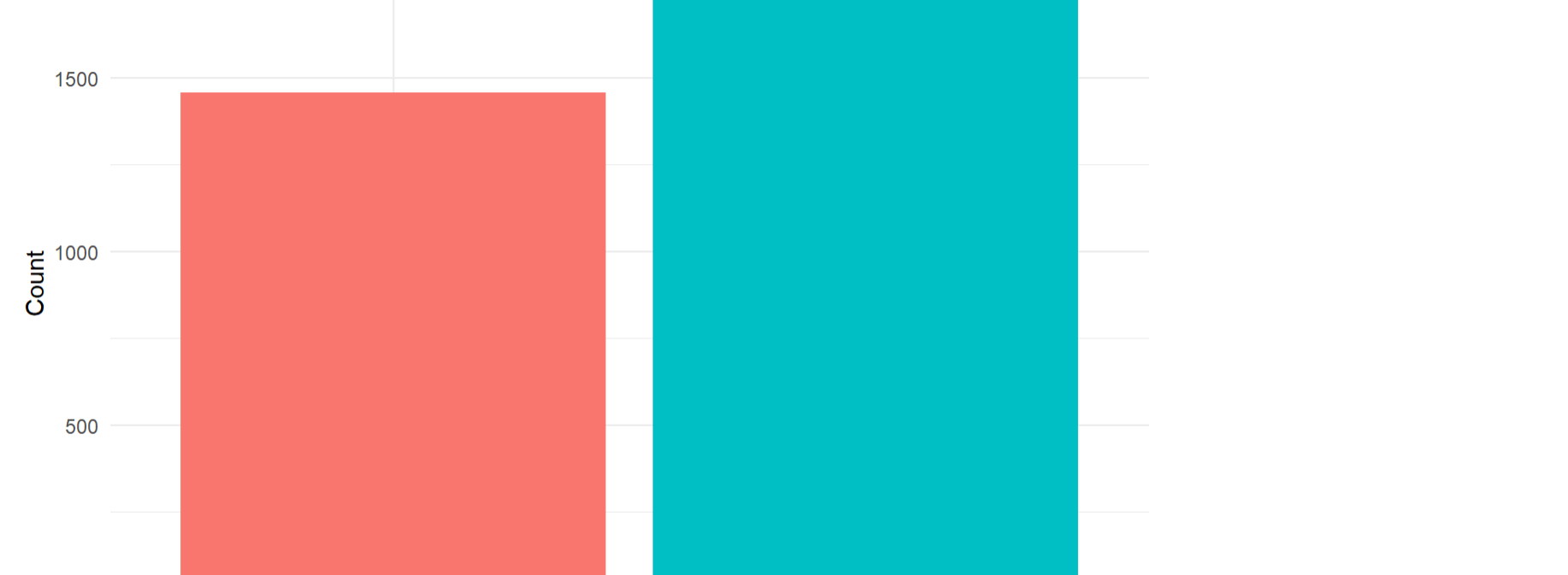
Digging deep into pH column

```
# Finding the number of samples having PH scale of Not portable water
ph.filter <- water.filtered %>%
  filter(ph < 6.5 | ph > 8.5)

ph.filter %>%
  summarise(Count = n())

## Count
## 1 1457

# Visualizing the count of pH scales within and outside the recommended limits.
water.filtered %>%
  mutate(ph.category = ifelse(ph < 6.5 | ph > 8.5, "Outside Limit", "Within Limit")) %>%
  group_by(ph.category) %>%
  summarise(Count = n()) %>%
  ggplot(aes(x = ph.category, y = Count, fill = ph.category)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme_minimal() +
  guides(fill = FALSE) +
  labs(title = "pH Scale Distribution: Within vs. Outside Recommended Limits")
```



```
# What is the average pH value across all water bodies?
mean(water.filtered$ph)

## [1] 7.080795
```

```
# What is the percentage of potable water bodies with a pH less than 7?
potable.water <- water.filtered %>%
  filter(Potability == "Potable")

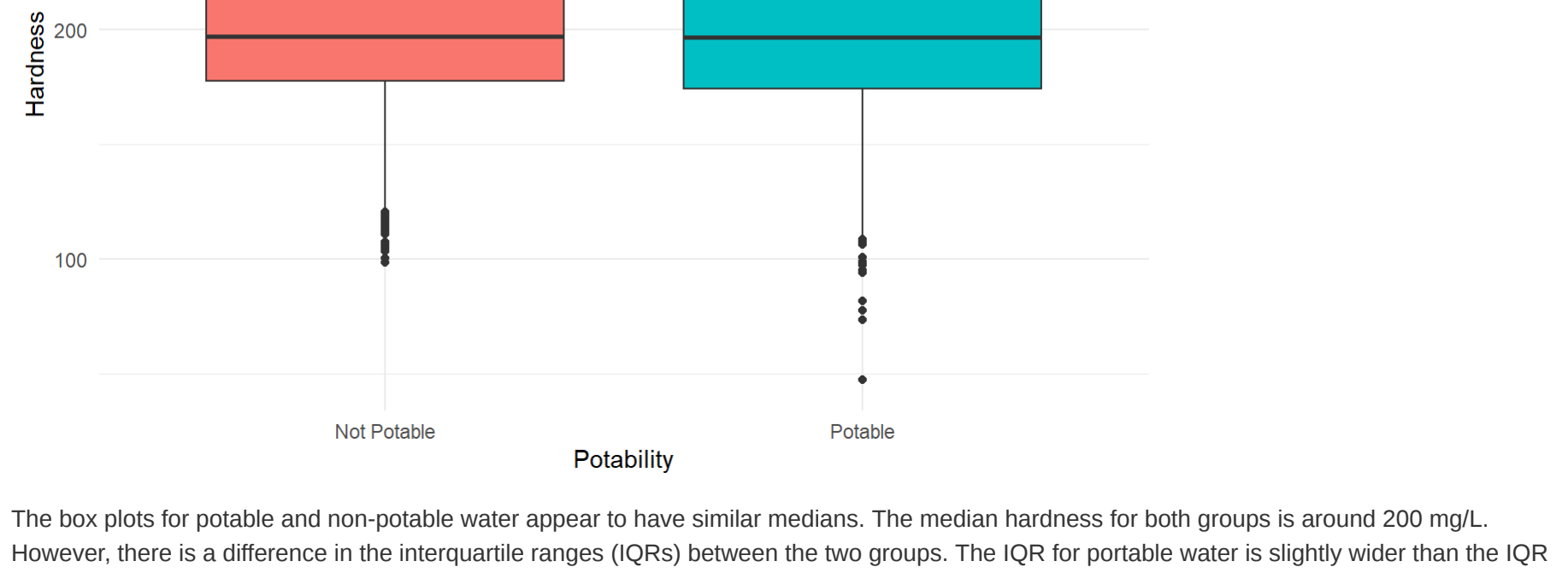
ph.7.less <- mean(potable.water$ph < 7) * 100

print(ph.7.less)

## [1] 42.09703
```

The ideal pH range for drinking water is between 6.5 and 8.5. This range is considered safe for human consumption. Despite a higher proportion of non-potable water in the sample, a majority of samples exhibit acceptable pH levels, with a count surpassing 1750. Based on this information, we can conclude that while there is a significant portion of non-potable water in the sample, a large number of samples still meet the pH standards for drinking water. This suggests that the pH level may not be the primary factor determining the potability of water in this sample.

```
# How does the distribution of hardness vary among potable and non-potable water?
water.filtered %>%
  ggplot(aes(x = Potability, y = Hardness, fill = Potability)) +
  geom_boxplot() + theme_minimal() +
  guides(fill = FALSE) +
  labs(title = "Distribution of Hardness for Potable and Non-Potable Water")
```



The box plots for potable and non-potable water appear to have similar medians. The median hardness for both groups is around 200 mg/L. However, there is a difference in the interquartile ranges (IQRs) between the two groups. The IQR for potable water is slightly wider than the IQR for non-potable water. The wider IQR for potable water suggests that the distribution of hardness for potable water is more spread out than the distribution of hardness for non-potable water. The wider IQR for potable water suggests that there are more extreme hardness values observed in potable water samples. This broader range of hardness levels could be due to various factors, such as differences in water treatment processes, natural variations in water sources, or contamination from external sources.

```
# What is the maximum and minimum total dissolved solids (TDS) recorded in potable water?
# method 1
summary(water.filtered$Solids)

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 320.9 15566.7 20927.8 22014.1 27332.8 61227.2
```

```
# method 2
max.value <- max(water.filtered$Solids)
min.value <- min(water.filtered$Solids)
print(max.value)

## [1] 61227.2

print(min.value)

## [1] 320.9426
```

```
# method 3, combining both results to one output
max.min.TDS <- rbind(max.value, min.value)
rownames(max.min.TDS) <- c("Max TDS:", "Min TDS:")
print(max.min.TDS)

##           [,1]
## Max TDS: 61227.1960
## Min TDS: 320.9426
```

```
# Is there any relationship between chloramines and sulfate concentrations in the water?
correlation <- cor(water.filtered$Chloramines, water.filtered$Sulfate)

print(correlation)

## [1] 0.02379109
```

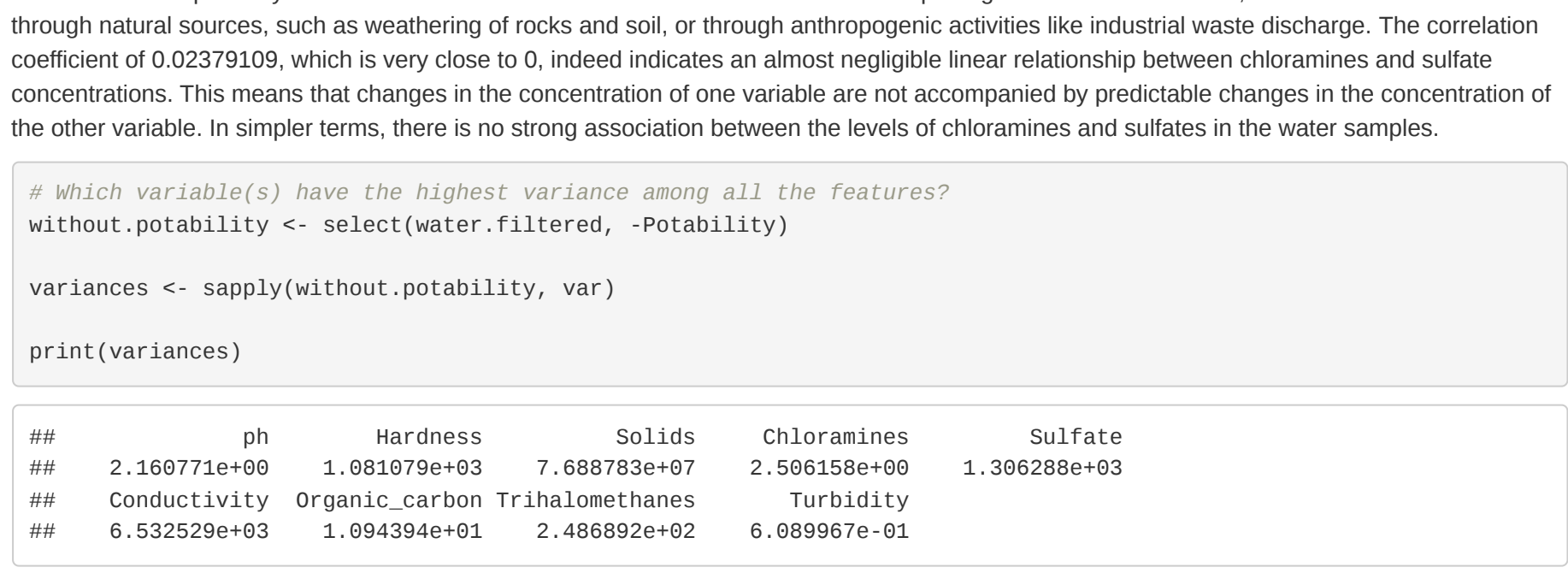
```
ggplot(water.filtered, aes(x = Chloramines, y = Sulfate)) +
  geom_point(alpha = 0.5) +
  geom_smooth() +
  labs(title = "Scatterplot between Chloramines and Sulfates")

## 'geom_smooth()' using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



```
# digging deeper
ggplot(water.filtered, aes(x = Chloramines, y = Sulfate,)) +
  geom_point(alpha = 0.5, color = "brown") +
  geom_smooth() +
  facet_wrap(~ Potability) +
  theme_minimal() +
  labs(title = "Scatterplot between Chloramines and Sulfates")

## 'geom_smooth()' using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



Chloramines are primarily added to water as a disinfectant to kill harmful bacteria and pathogens. On the other hand, sulfates can enter water through natural sources, such as weathering of rocks and soil, or through anthropogenic activities like industrial waste discharge. The correlation coefficient of 0.02379109, which is very close to 0, indeed indicates an almost negligible linear relationship between chloramines and sulfate concentrations. This means that changes in the concentration of one variable are not accompanied by predictable changes in the concentration of the other variable. In simpler terms, there is no strong association between the levels of chloramines and sulfates in the water samples.

```
# Which variable(s) have the highest variance among all the features?
without.potability <- select(water.filtered, ~Potability)
variances <- sapply(without.potability, var)
print(variances)

##           ph           Hardness           Solids           Chloramines           Sulfate
##  2.169771e+08  1.083079e+03  7.688783e+07  2.506158e+06  1.396288e+03
##  6.525259e+03  1.094394e+01  2.486892e+02  6.089997e-01
```

Solids are having the highest variance, with a variance of 7.688783e+07

Variances provide insights into the distribution and spread of data within each variable. They help assess the degree of variability, which is crucial for understanding the characteristics and nature of each feature in our dataset. Solids, Conductivity and Turbidity have more significant variability or dispersion in their data points compared to other variables. Higher variance suggests that data points are more spread out from the mean, indicating a wider range of values in the dataset.

```
# How many water bodies have sulfate levels above the recommended limit for drinking purposes?
sulfate.limit <- water.filtered %>%
  filter(Sulfate > 250) # The permissible limit of sulfate level in drinking water is 250 mg/L

length(sulfate.limit$Sulfate)

## [1] 3218
```

```
# Is there a significant difference in the mean chloramine levels between potable and non-potable water?
potable.water <- water.filtered %>%
  filter(Potability == "Potable")

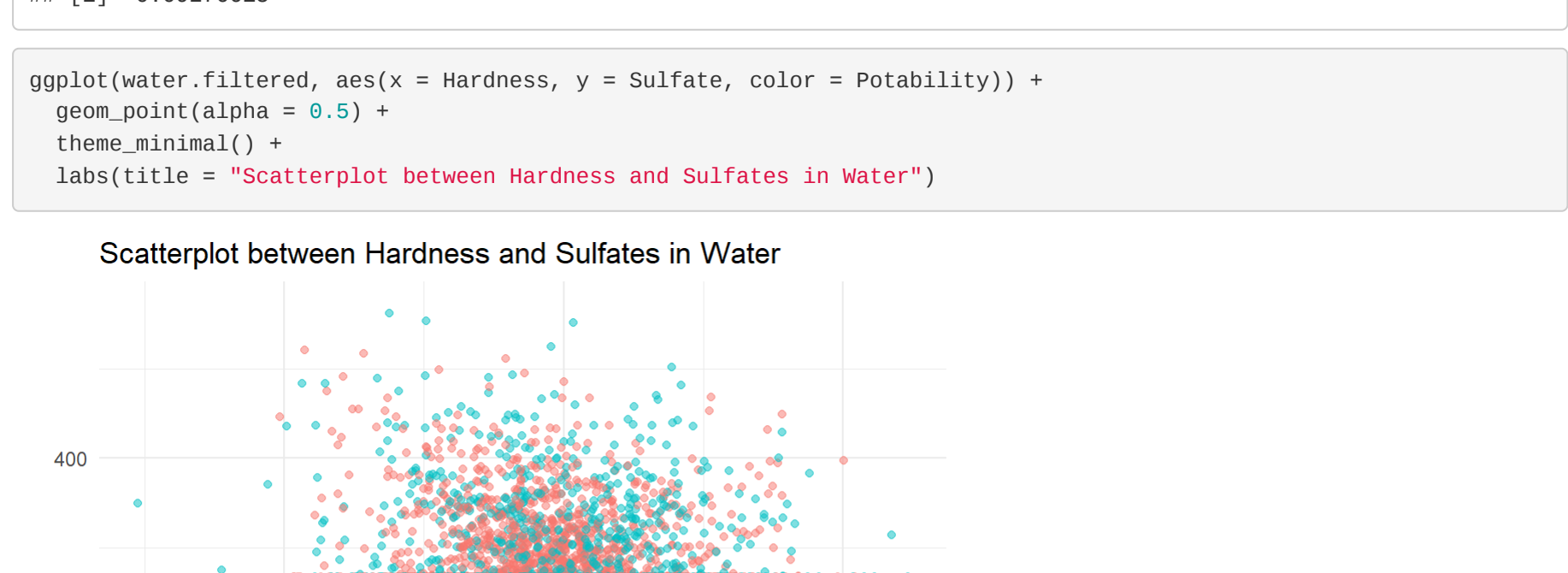
non.potable <- water.filtered %>%
  filter(Potability == "Not Potable")

# Perform t-test to compare mean Chloramine levels between potable and non-potable water
t.test.result <- t.test(potable.water$Chloramines, non.potable$Chloramines)

t.test.result

##
## Welch Two Sample t-test
##
## data:  potable.water$Chloramines and non.potable$Chloramines
## t = 1.3239, df = 2471.3, p-value = 0.1856
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.0272659 0.19145351
## sample estimates:
## mean of x mean of y
##  7.169338  7.092175
```

```
ggplot(water.filtered, aes(x = factor(Potability), y = Chloramines, fill = Potability)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Chloramine Levels between Potable and Non-potable Water", x = "Potability") +
  guides(fill = FALSE)
```



The box plot shows that the median chloramine level in potable water is higher than the median chloramine level in non-potable water. However, the interquartile range (IQR) for potable water is also wider, indicating that the distribution of chloramine levels in potable water is more spread out. Based on the results of this Welch Two Sample t-test, with a p-value of 0.1856 greater than the typical significance level of 0.05, there isn't enough evidence to conclude a significant difference between the means of Chloramines in potable and non-potable water.

```
# What is the relationship between hardness and sulfate levels in the water?
cor.value <- cor(water.filtered$Hardness, water.filtered$Sulfate)

print(cor.value)

## [1] -0.09276615
```

```
ggplot(water.filtered, aes(x = Hardness, y = Sulfate, color = Potability)) +
  geom_point(alpha = 0.5) +
  theme_minimal() +
  labs(title = "Scatterplot between Hardness and Sulfates in Water")

# guides(fill = FALSE)
```

```
## $fill
## [1] "none"
## attr(,"class")
## [1] "guides"
```

The obtained correlation value of approximately -0.0928 indicates a weak negative correlation between hardness and sulfates in water. This means that as the hardness of water increases, the concentration of sulfates tends to decrease slightly. However, it is important to note that the correlation coefficient is very close to 0, which suggests that the relationship between the two variables is very weak.

Data Analysis and Insights:

- pH Levels:**
 - Majority of samples fell within the WHO recommended pH range of 6.5 to 8.5 for safe drinking water.
 - Approximately 42% of potable water bodies had a pH less than 7, still within acceptable limits.
- Hardness and Potability:**
 - Similar median hardness observed for potable and non-potable water, but wider range of values in potable water suggests more extreme hardness levels.
- Total Dissolved Solids (TDS):**
 - TDS levels varied considerably, with the highest variance among the features.
 - Several water bodies recorded TDS values exceeding the recommended limit for drinking purposes.
- Chloramines and Sulfate:**
 - Negligible linear relationship found between chloramines and sulfate concentrations.
 - Potable water showed a slightly higher median chloramine level but with a wider spread.
- Correlation between Hardness and Sulfate:**
 - A weak negative correlation exists between hardness and sulfate levels.

Key Takeaways:

- pH Compliance:** A majority of samples met pH standards despite some deviation, implying satisfactory acid-base balance.
- TDS and Potability:** Elevated TDS levels observed in a considerable number of water bodies could raise concerns regarding drinkability.
- Chloramines Comparison:** No significant difference in mean chloramine levels between potable and non-potable water, though potable water showed a wider spread.
- Correlation:** Weak correlations imply minimal direct relationships between some parameters, highlighting independent influences on water quality.

Conclusion:

The assessment provides insights into various water quality parameters, indicating areas of compliance and concern.

Further investigation is recommended for TDS levels surpassing drinking water limits, while understanding the nuances of factors influencing water quality is crucial for effective policymaking and ensuring safe drinking water for communities.