```
library(ggplot2)
```

```
## Warning: package ’ggplot2’ was built under R version 4.1.2
```

```
library(dplyr)
```

```
## Warning: package ’dplyr’ was built under R version 4.1.2
```

```
## 
## Attaching package: ’dplyr’
```

```
## The following objects are masked from ’package:stats’:
## 
##     filter, lag
```

```
## The following objects are masked from ’package:base’:
## 
##     intersect, setdiff, setequal, union
```

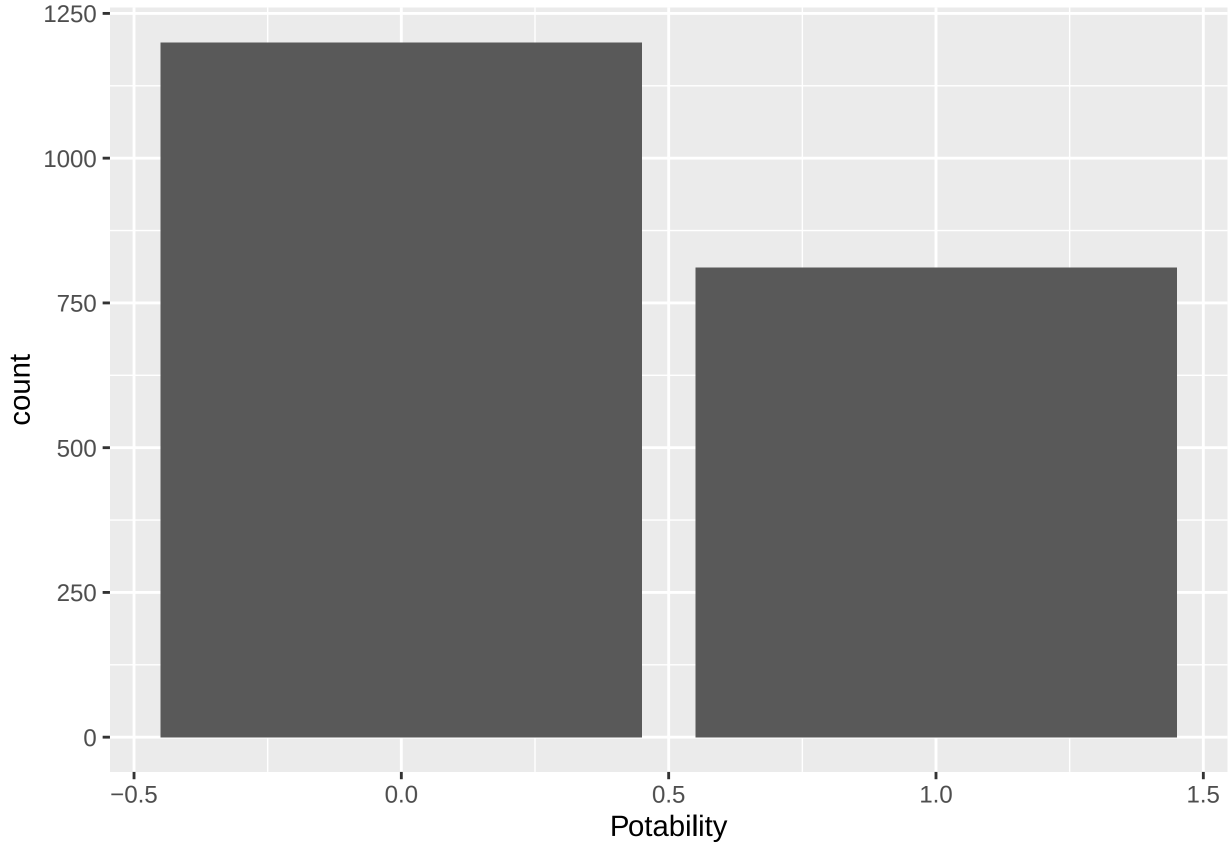
```
#####data import and cleaning #####
raw_data <- read.csv("C:/Users/DEV/Downloads/Documents/Data Science/Water Quality Anaylsis/water_potabil

summary(raw_data)
```

```
##        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    
##  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  
## 
```

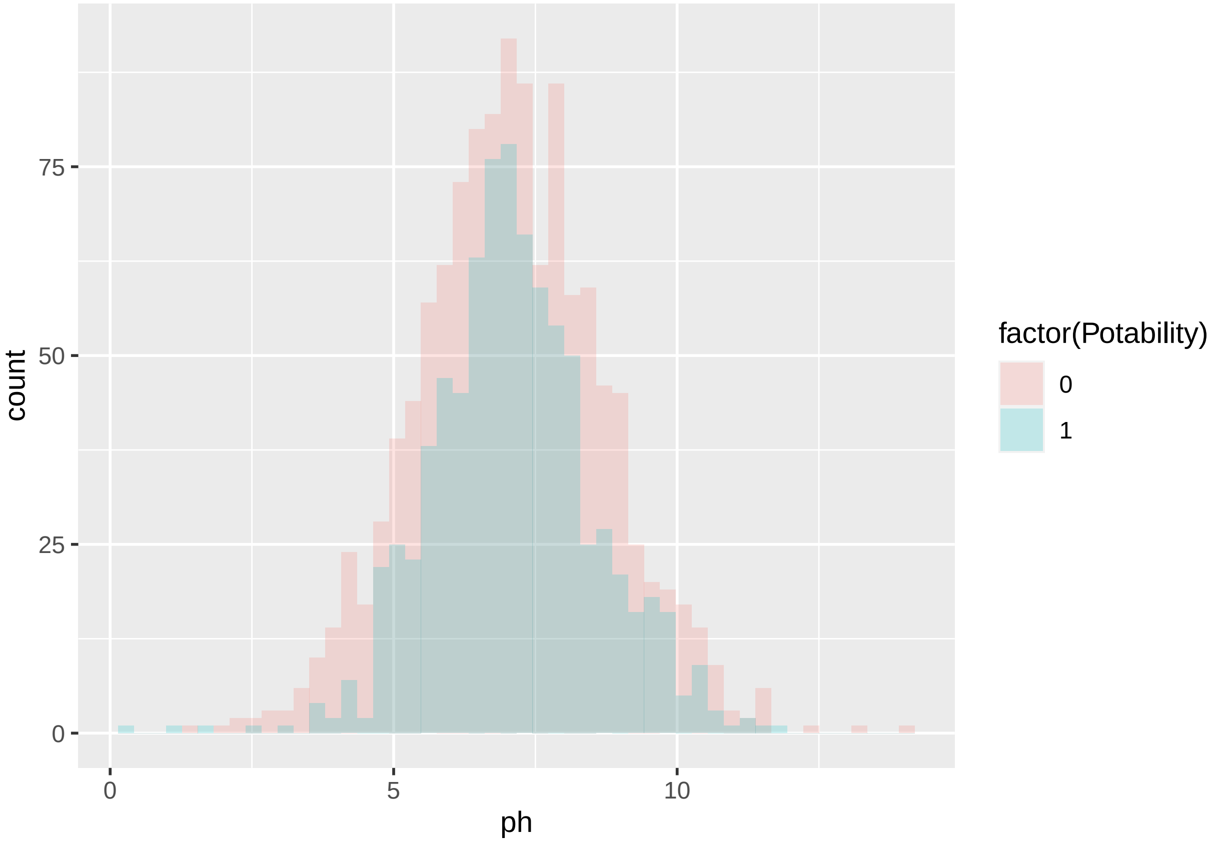
```
colnames(raw_data)
```

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

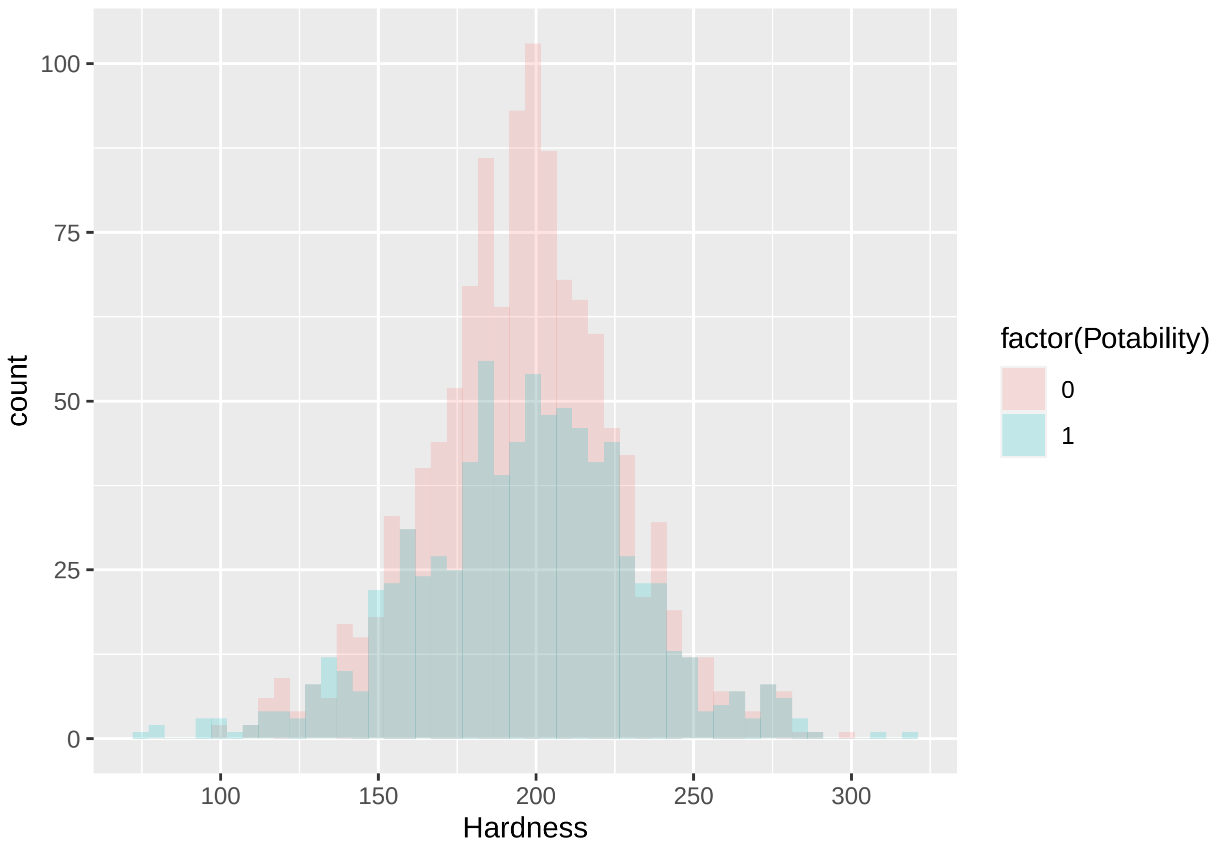
```
##removing NAs

data <- raw_data[complete.cases(raw_data),]

## imbalance of Potability
ggplot2::ggplot(data,aes(Potability))+
  ggplot2::geom_bar()
```



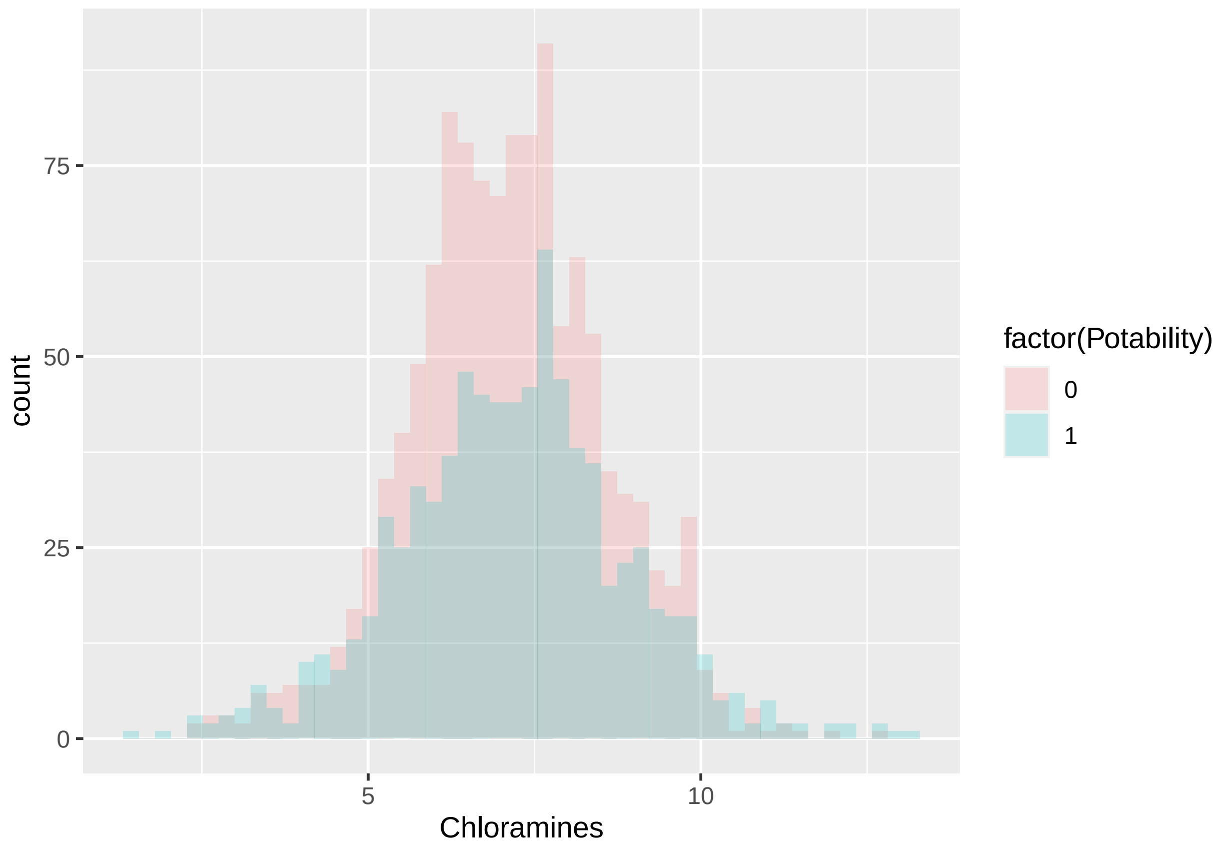
```
##
ggplot(data,aes(ph , fill  = factor(Potability)))+
  geom_histogram(position = "identity", alpha = 0.2, bins = 50)
```



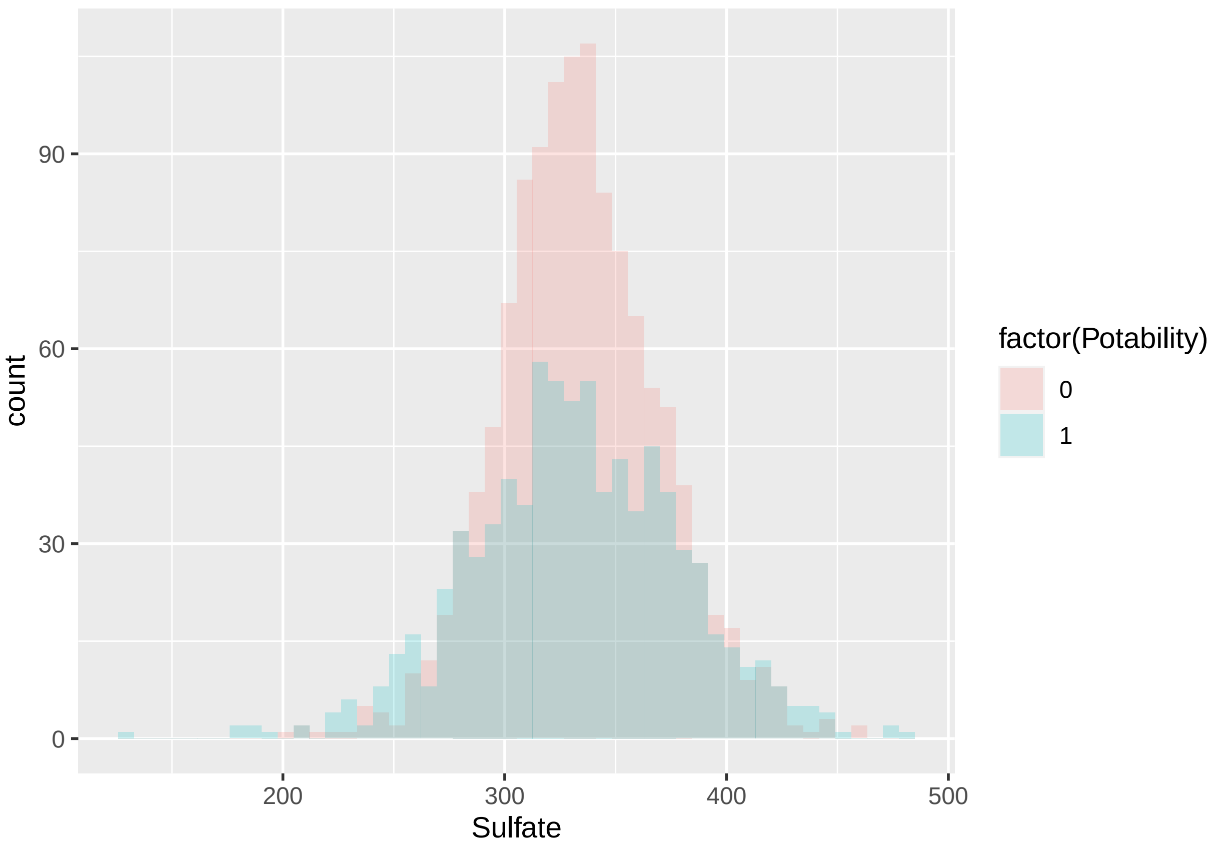
```
##
ggplot(data,aes( Hardness, fill  = factor(Potability)))+
  geom_histogram(position = "identity", alpha = 0.2, bins = 50)
```



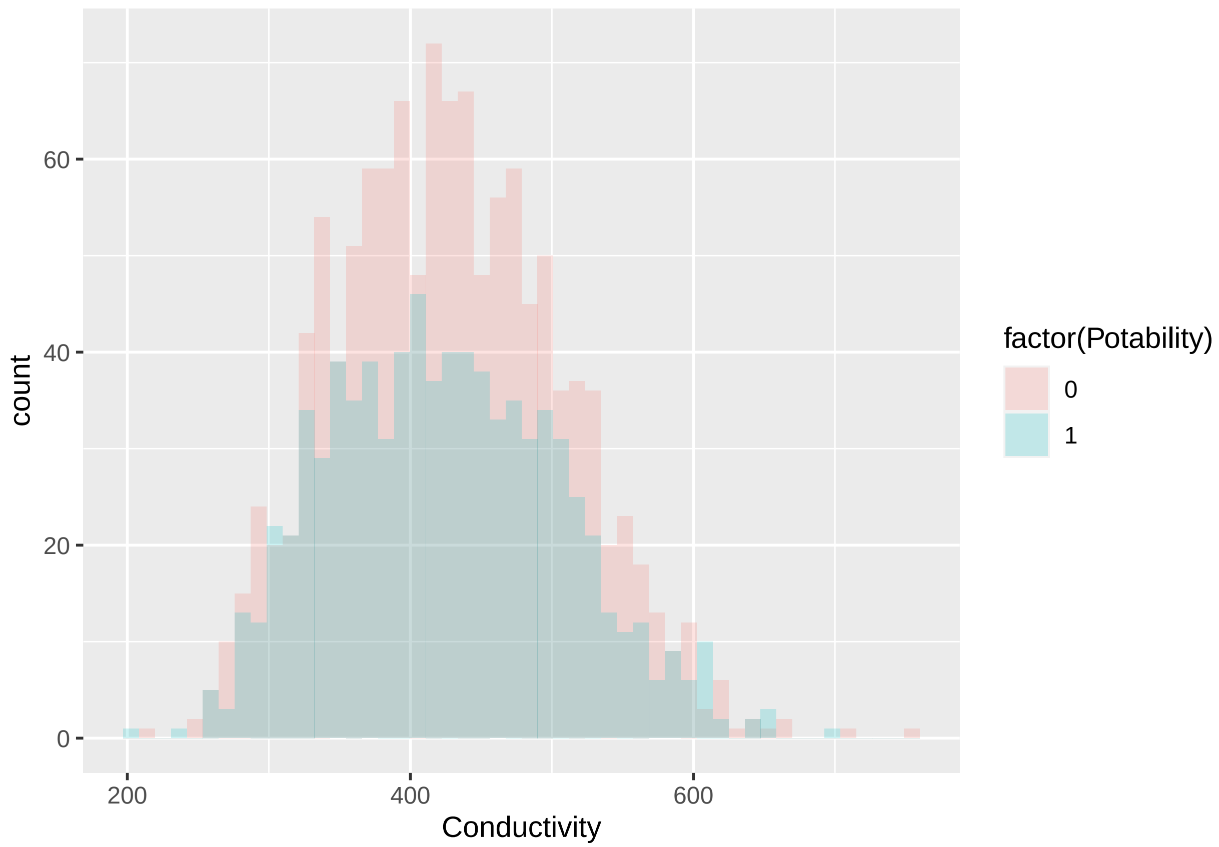
```
##
ggplot(data,aes(Chloramines , fill  = factor(Potability)))+
  geom_histogram(position = "identity", alpha = 0.2, bins = 50)
```



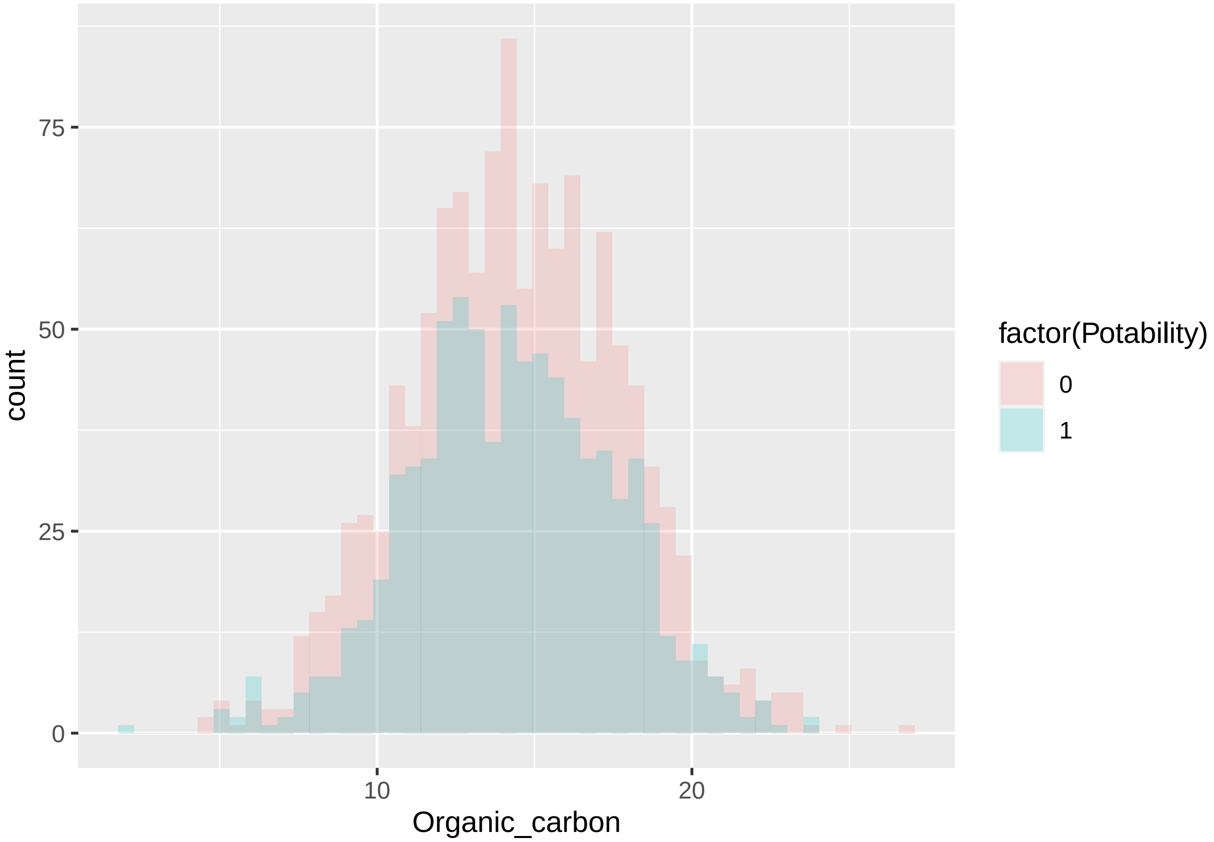
```
##
ggplot(data,aes(Sulfate, fill  = factor(Potability)))+
  geom_histogram(position = "identity", alpha = 0.2, bins = 50)
```



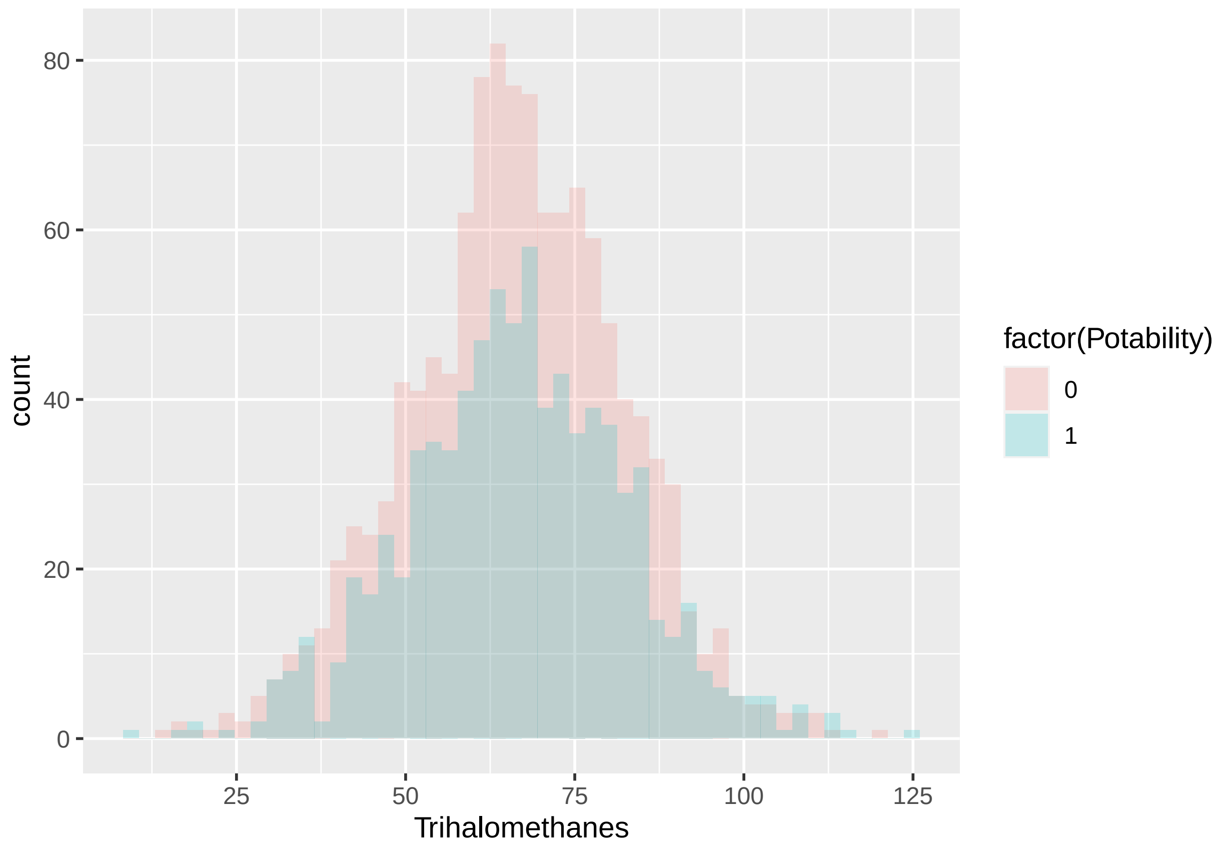
```
##
ggplot(data,aes( Conductivity, fill  = factor(Potability))+
  geom_histogram(position = "identity", alpha = 0.2, bins = 50)
```



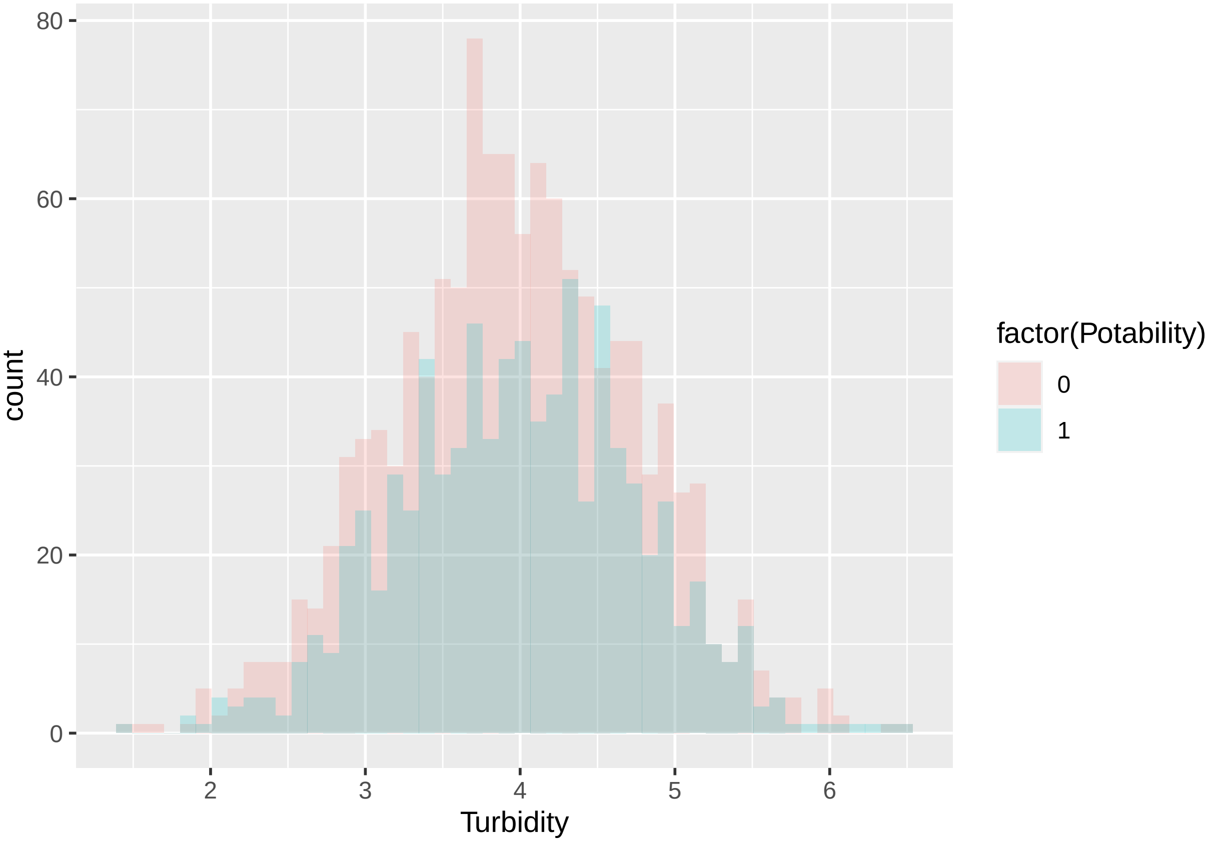
```
##
ggplot(data,aes( Organic_carbon, fill  = factor(Potability)))+
  geom_histogram(position = "identity", alpha = 0.2, bins = 50)
```



```
##
ggplot(data,aes( Trihalomethanes, fill  = factor(Potability)))+
  geom_histogram(position = "identity", alpha = 0.2, bins = 50)
```



```
##
ggplot(data,aes( Turbidity, fill  = factor(Potability)))+
  geom_histogram(position = "identity", alpha = 0.2, bins = 50)
```


80
60
40
20
0
count
2
3
4
5
6
Turbidity
factor(Potability)
0
1