DATA624_HW4

John Ferrara

2025-03-01

Homework 4

Do problems 3.1 and 3.2 in the Kuhn and Johnson book Applied Predictive Modeling. Please submit your Rpubs link along with your .pdf for your run code.

3.1.

The UC Irvine Machine Learning Repository6 contains a data set related to glass identification. The data consist of 214 glass samples labeled as one of seven class categories. There are nine predictors, including the refractive index and percentages of eight elements: Na, Mg, Al, Si, K, Ca, Ba, and Fe. The data can be accessed via:

```
data(Glass)
print(str(Glass))
```

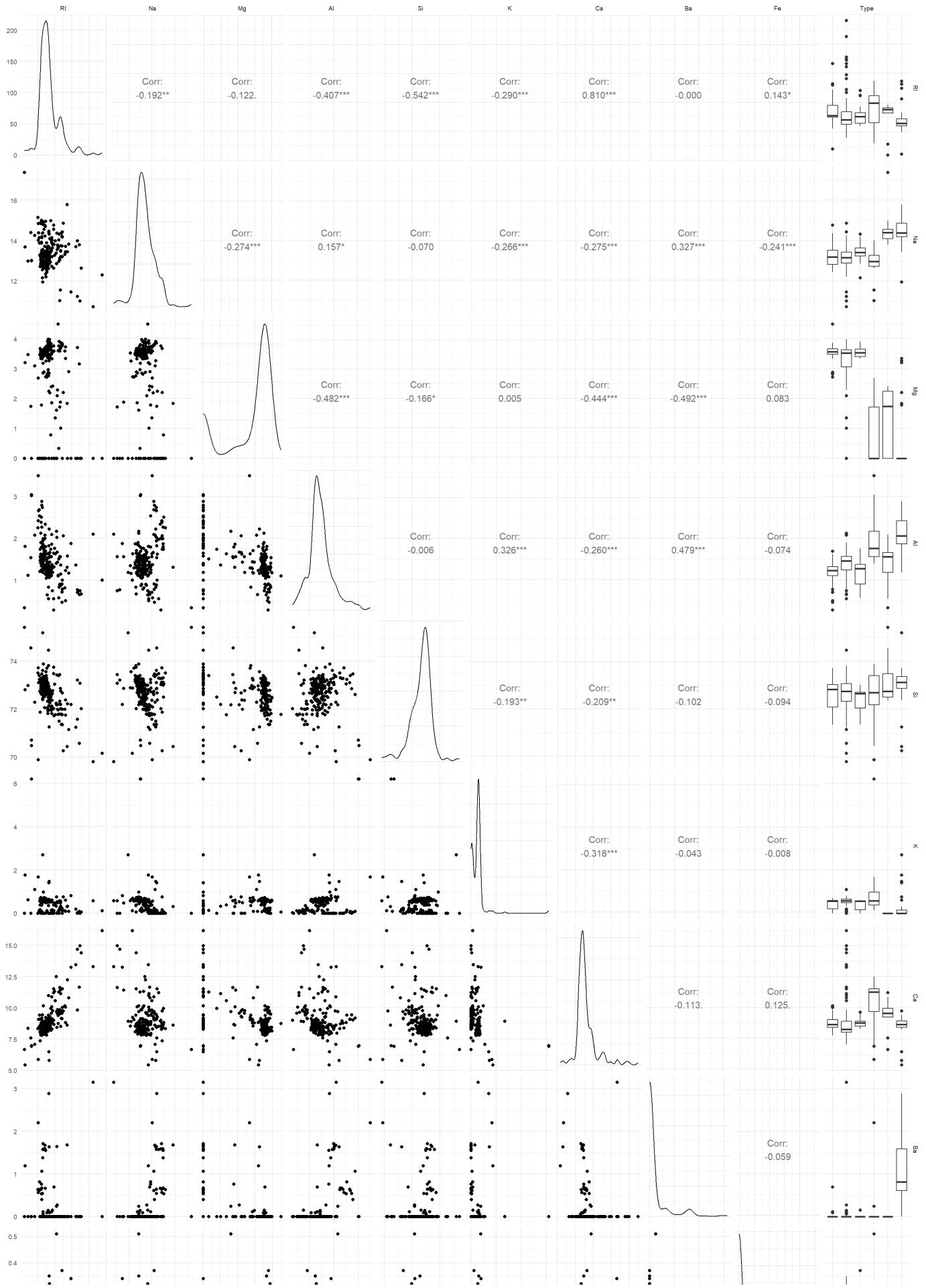
```
## 'data.frame':    214 obs. of  10 variables:
##  $ RI  : num  1.52 1.52 1.52 1.52 1.52 ...
##  $ Na  : num  13.6 13.9 13.5 13.2 13.3 ...
##  $ Mg  : num  4.49 3.6 3.55 3.69 3.62 3.61 3.6 3.61 3.58 3.6 ...
##  $ Al  : num  1.1 1.36 1.54 1.29 1.24 1.62 1.14 1.05 1.37 1.36 ...
##  $ Si  : num  71.8 72.7 73 72.6 73.1 ...
##  $ K   : num  0.06 0.48 0.39 0.57 0.55 0.64 0.58 0.57 0.56 0.57 ...
##  $ Ca  : num  8.75 7.83 7.78 8.22 8.07 8.07 8.17 8.24 8.3 8.4 ...
##  $ Ba  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Fe  : num  0 0 0 0 0 0.26 0 0 0 0.11 ...
##  $ Type: Factor w/ 6 levels "1","2","3","5",..: 1 1 1 1 1 1 1 1 1 1 ...
## NULL
```

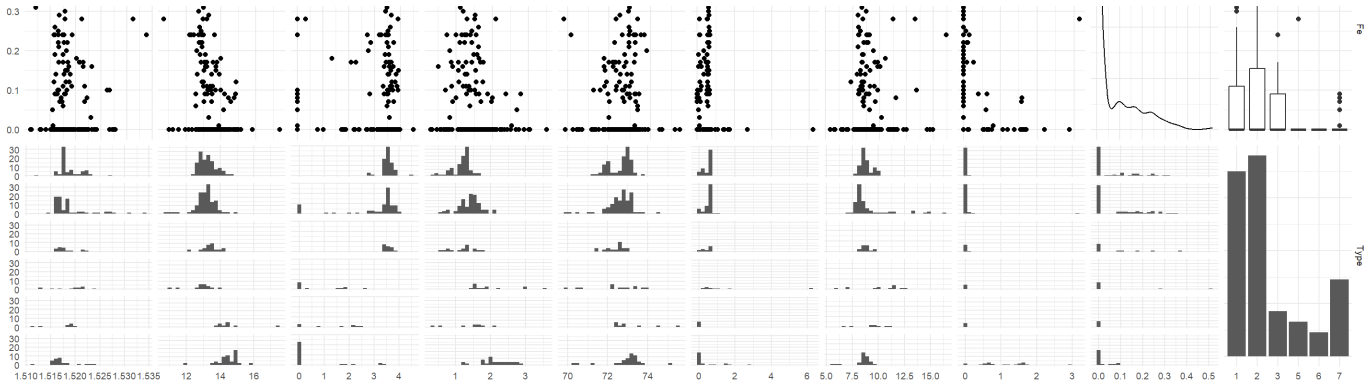
(a) Using visualizations, explore the predictor variables to understand their distributions as well as the relationships between predictors.

Based on the plots made by GGpairs, there are multiple noteworthy relationships within the data. Firstly, the only distribution of the the columns that seems to be close to normal is that of Si, while most of the other columns and values seem to have right skewness. That is with the exception of Mg, which is left skewed. With respect to the relationships between variables, those that are visually of note are: - Cs and the Reflective index have a pretty strong direct relationship based on the scatterplot of both columns. - Similarly, the reflective index (RI) and Na have a direct linear relationship as well, although not as pronounced. - Na also seems to have a direct linear relationship with Al in the data. - The RI may have a negative correlation with Al based on the looks of the scatterplot. - Bs and Na may have a positive non-linear, perhaps exponential relationship. - There may be a slight negative linear relationship with Mg and Ca.

```
## Using GGpairs to plot everythign for the varibles
ggpairs(Glass, progress = FALSE) +  theme_minimal(base_size=9)
```

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


RI
Na
Mg
Al
Si
K
Ca
Ba
Fe
Type
Corr: -0.192**
Corr: -0.122.
Corr: -0.407***
Corr: -0.542***
Corr: -0.290***
Corr: 0.810***
Corr: -0.000
Corr: 0.143*
Corr: -0.274***
Corr: 0.157*
Corr: -0.070
Corr: -0.266***
Corr: -0.275***
Corr: 0.327***
Corr: -0.241***
Corr: -0.482***
Corr: -0.166*
Corr: 0.005
Corr: -0.444***
Corr: -0.492***
Corr: 0.083
Corr: -0.006
Corr: 0.326***
Corr: -0.260***
Corr: 0.479***
Corr: -0.074
Corr: -0.193**
Corr: -0.209**
Corr: -0.102
Corr: -0.094
Corr: -0.318***
Corr: -0.043
Corr: -0.008
Corr: -0.113.
Corr: 0.125.
Corr: -0.059

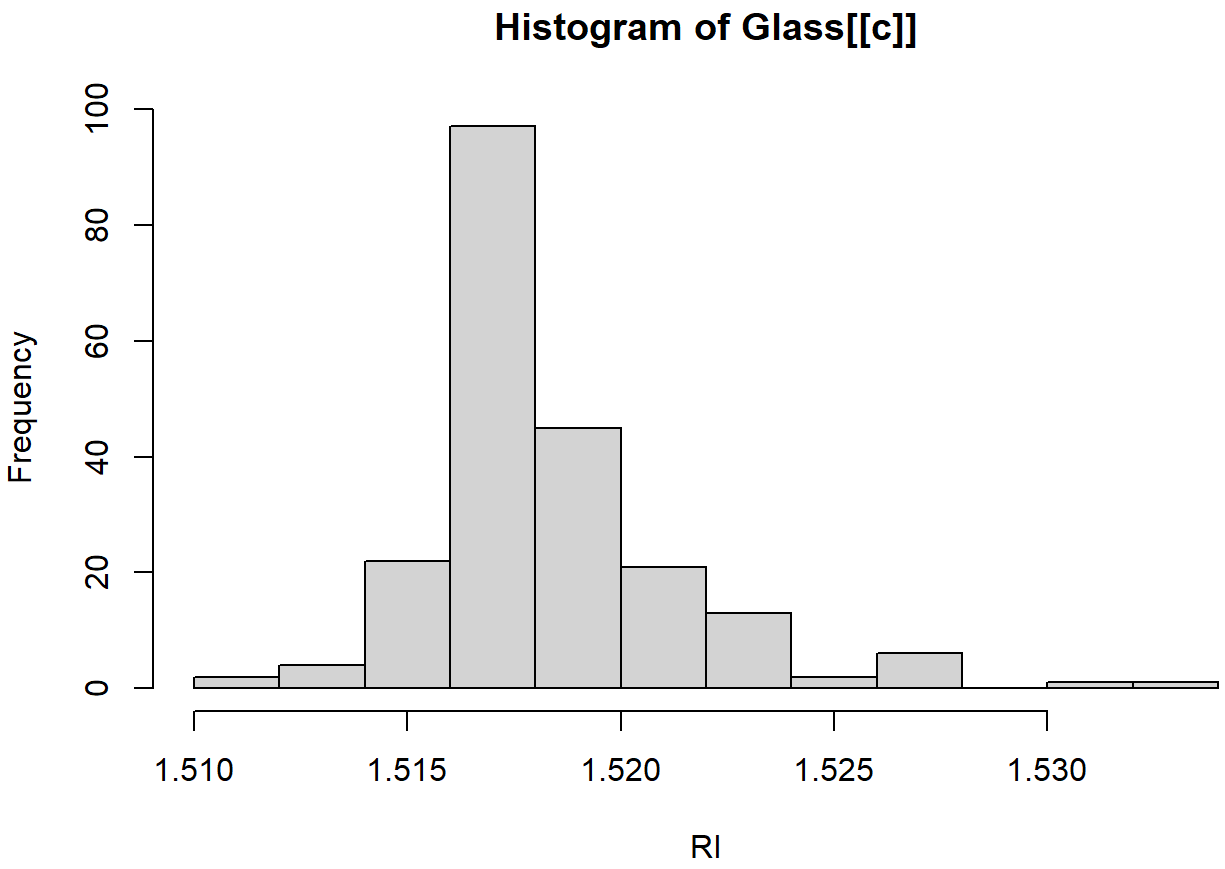


(b) Do there appear to be any outliers in the data? Are any predictors skewed?

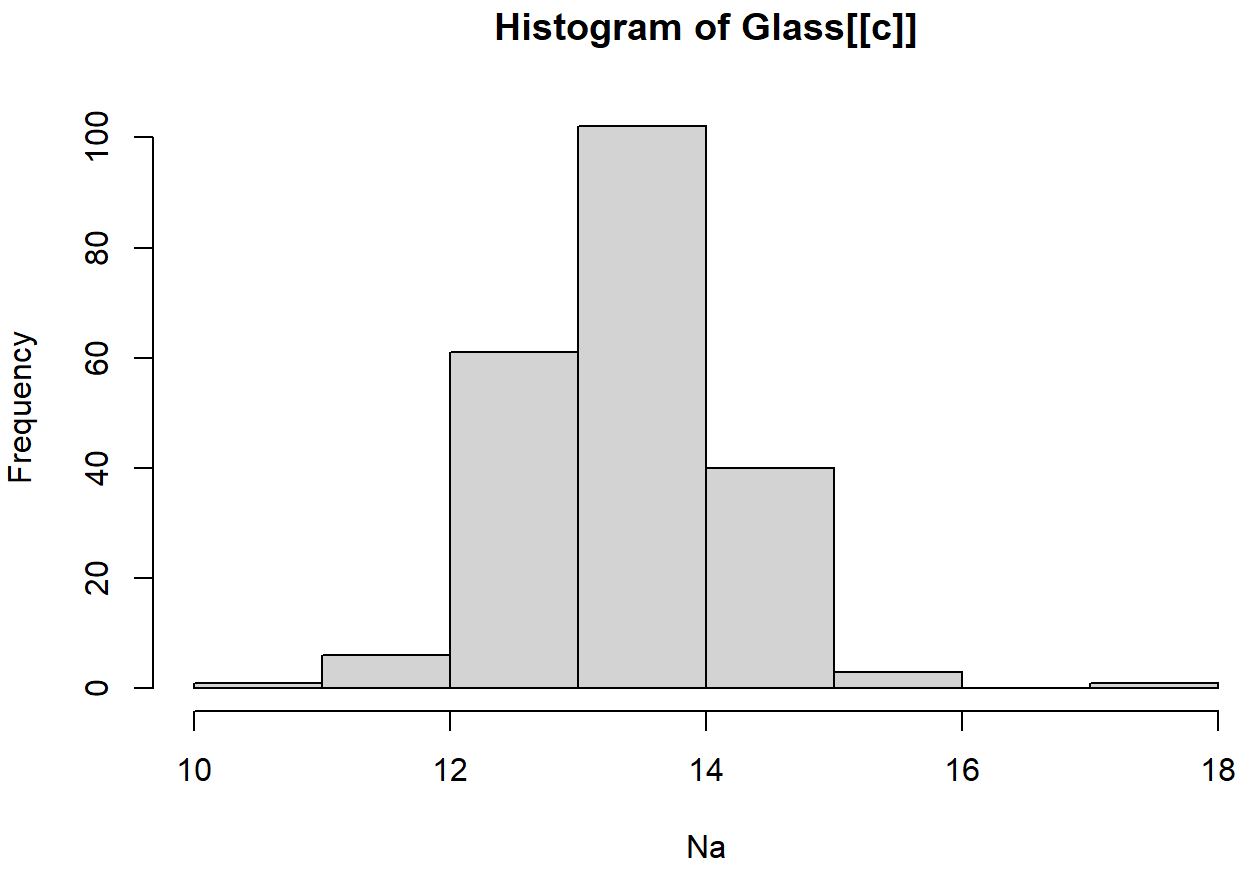
Based on the ggpairs plot and the custom histogram and skewness values below there is are skewed predictors here. The following predictors have right skewed data: RI, K, Ca, Ba, Fe. The following predictors have left skewed data: Mg, and Si. THose columns that have skewness but it may be a bit varied, or not too visible, are: Al. Additionally, based on the histograms the predictors that seem to have outliers are: RI, K, Ca, Ba and Fe.

```
## Plotting direct histograms for this data to look at skewness.

for (c in colnames(Glass)) {
  if (c == 'Type'){
    print("Type column is not numeric")
    NULL
  }
  else{
  print(c)
  hist(Glass[[c]],xlab = c)

  }
}
```

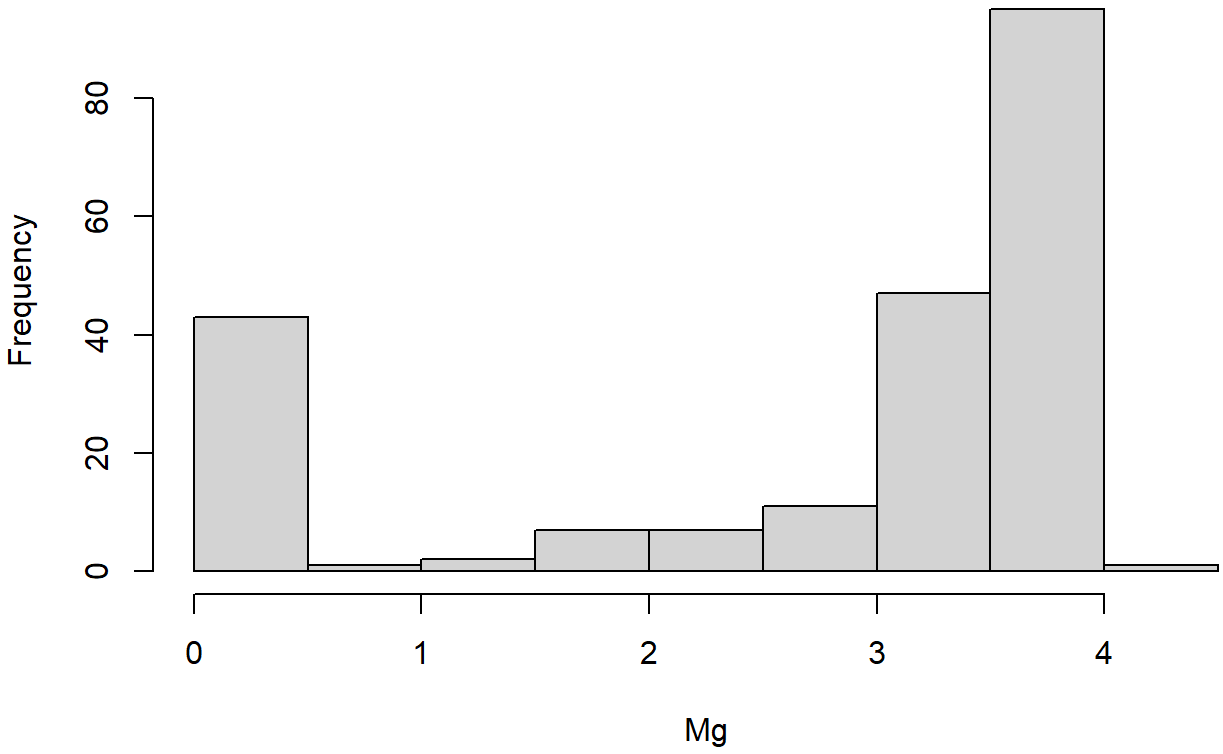
```
## [1] "RI"
```


Histogram of Glass[[c]]
Frequency
0
20
40
60
80
100
1.510
1.515
1.520
1.525
1.530
RI

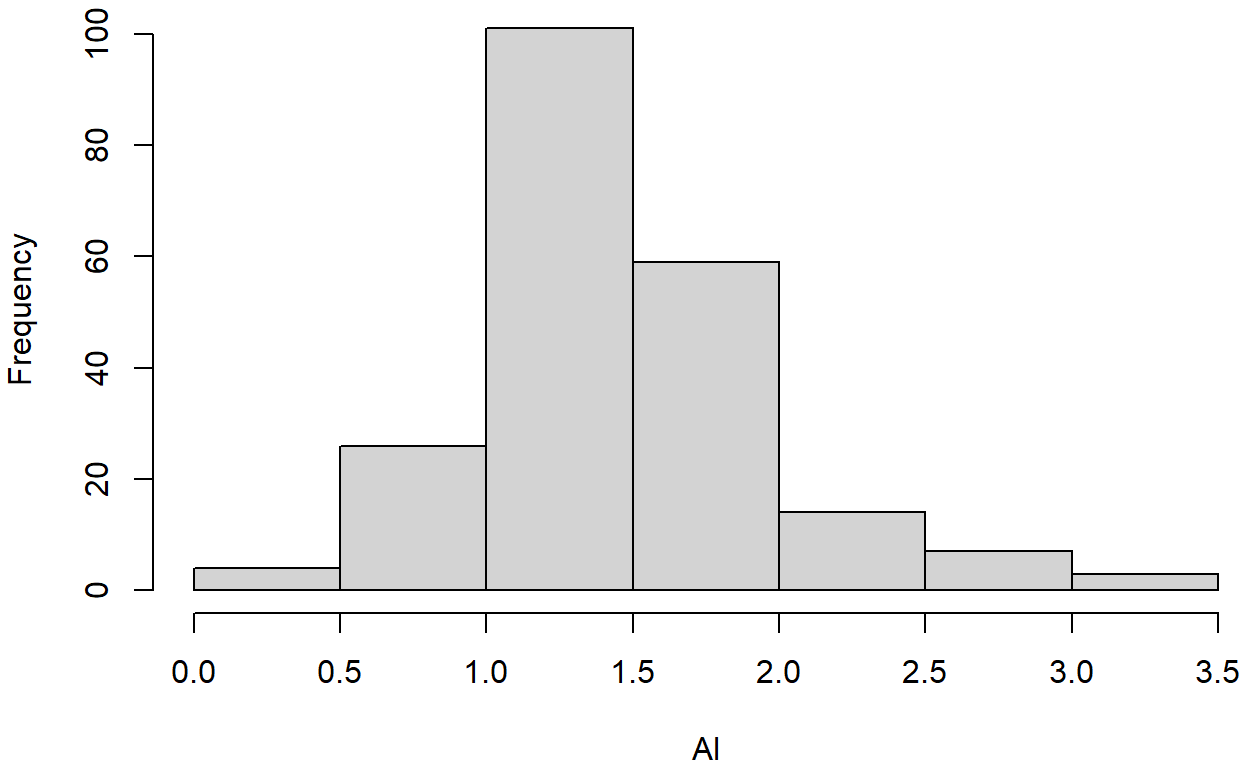
```
## [1] "Na"
```


Histogram of Glass[[c]]
Frequency
0
20
40
60
80
100
10
12
14
16
18
Na

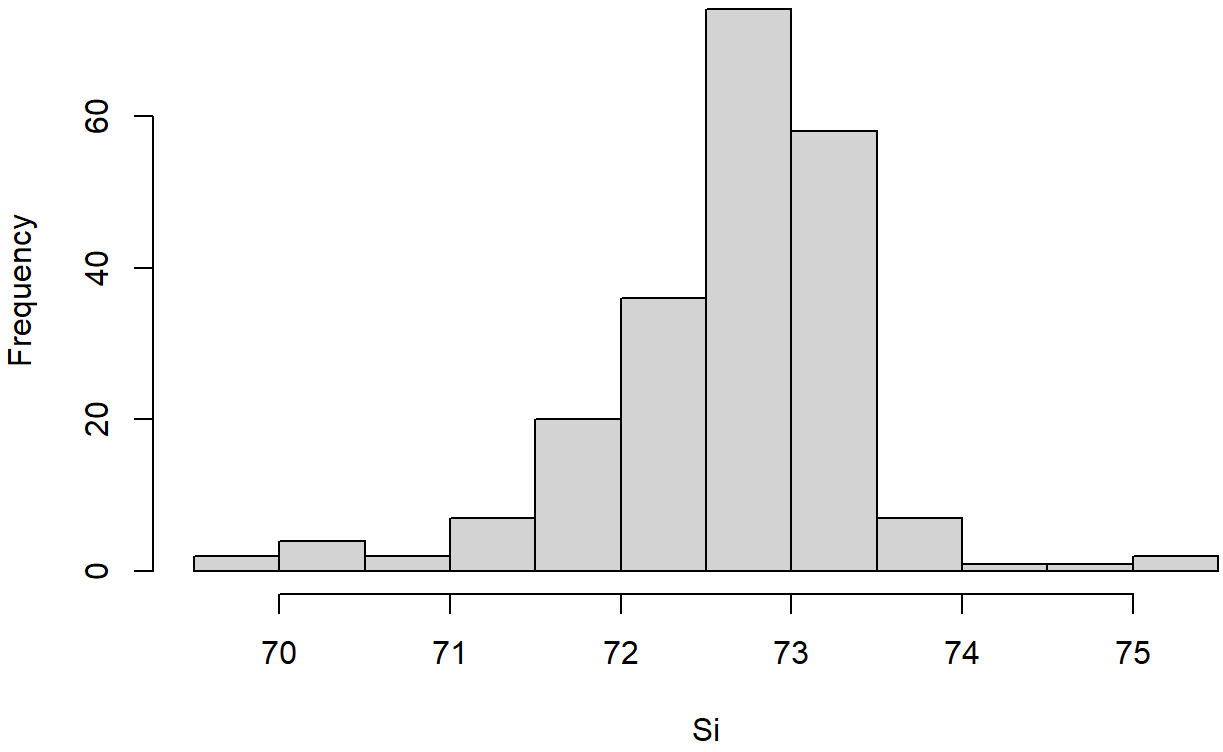
```
## [1] "Mg"
```



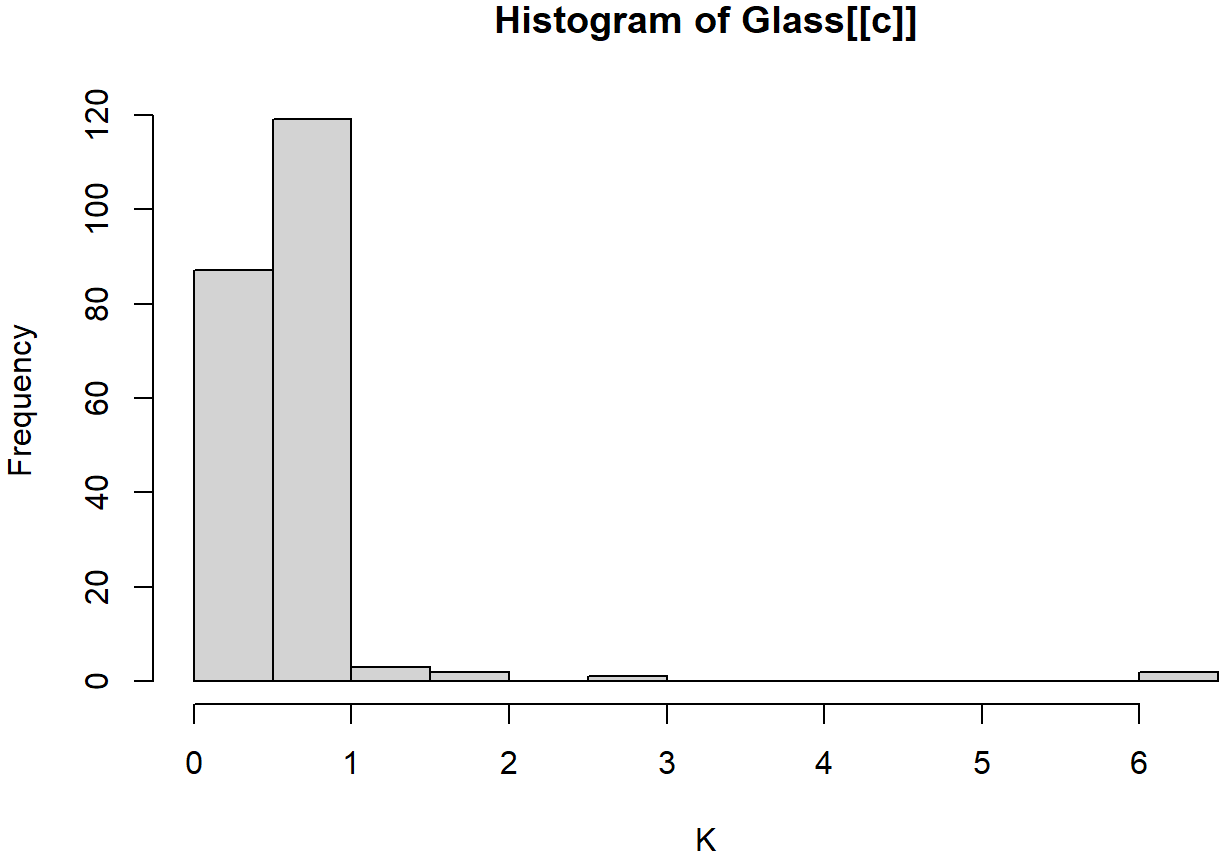

```
## [1] "Al"
```


Histogram of Glass[[c]]
Frequency
0
20
40
60
80
100
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Al

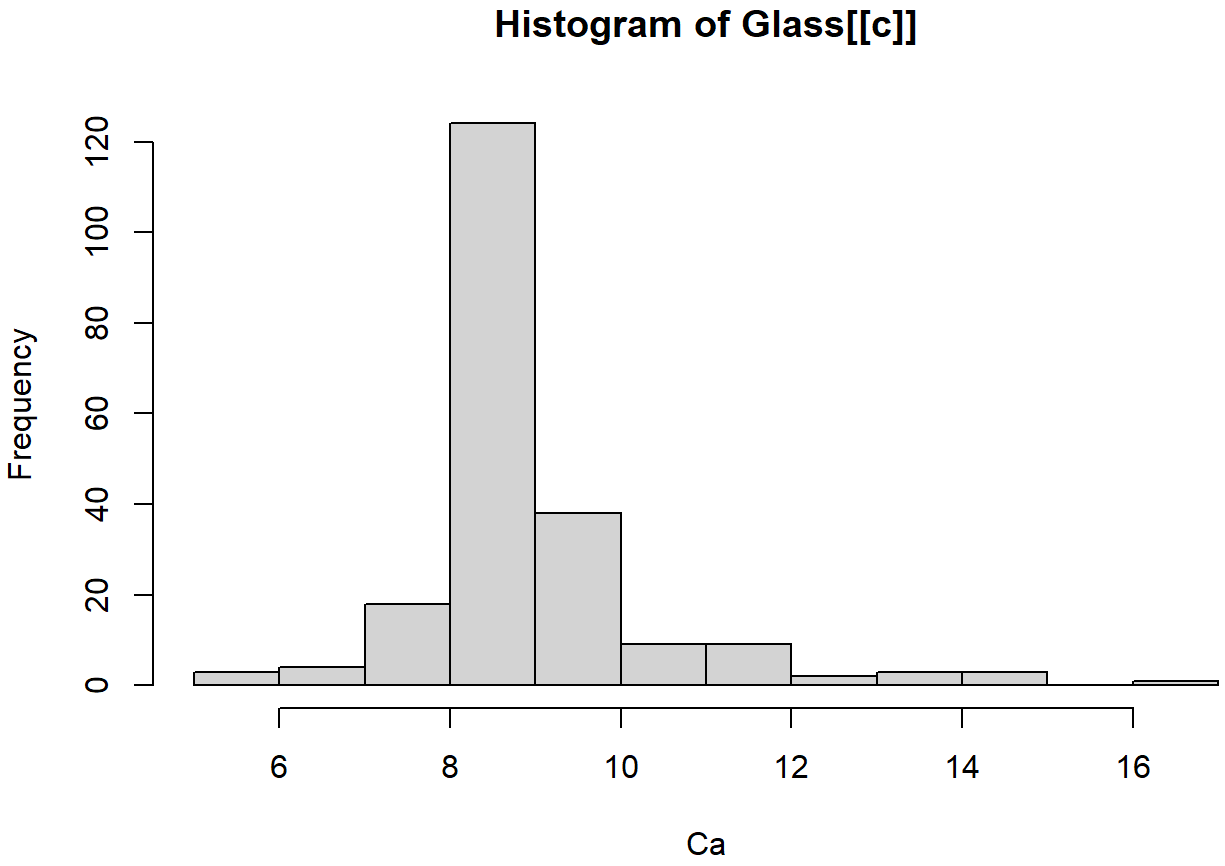
```
## [1] "Si"
```



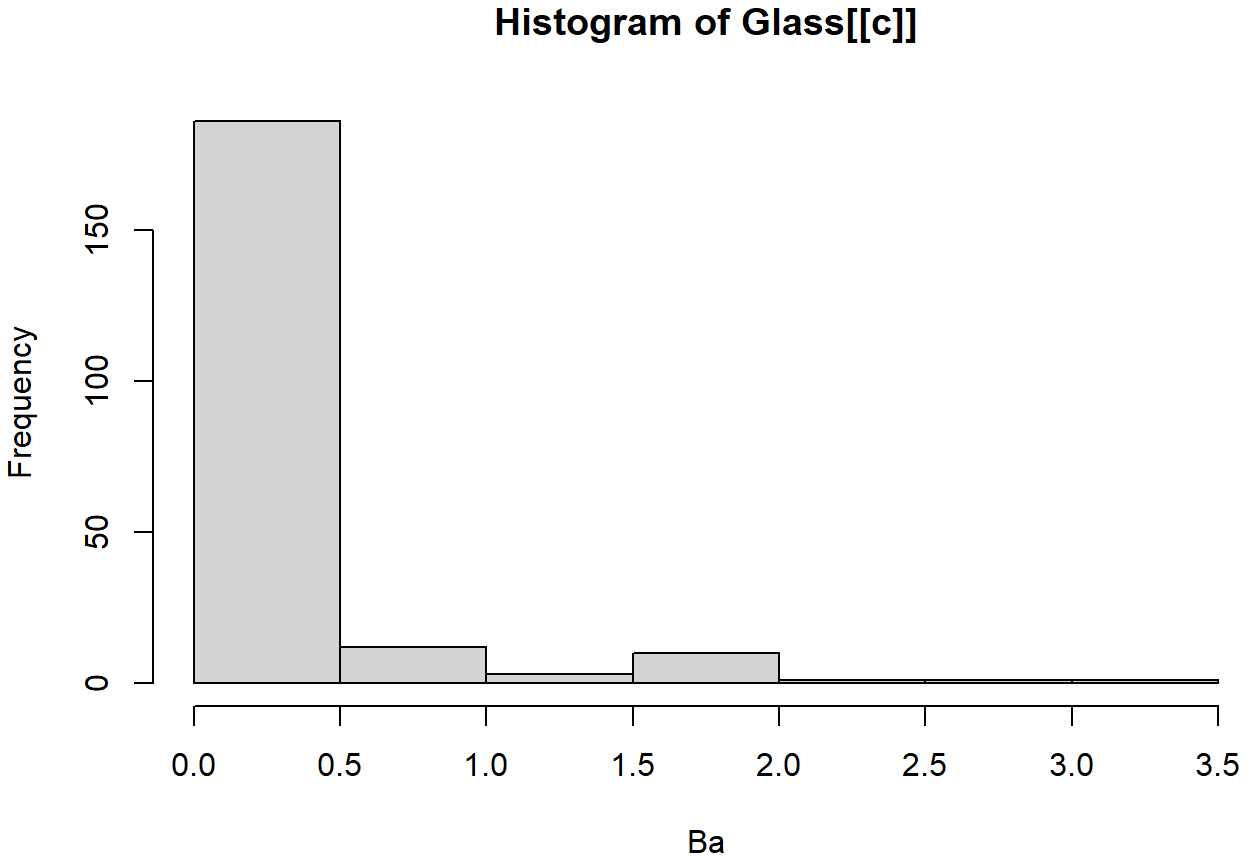

```
## [1] "K"
```



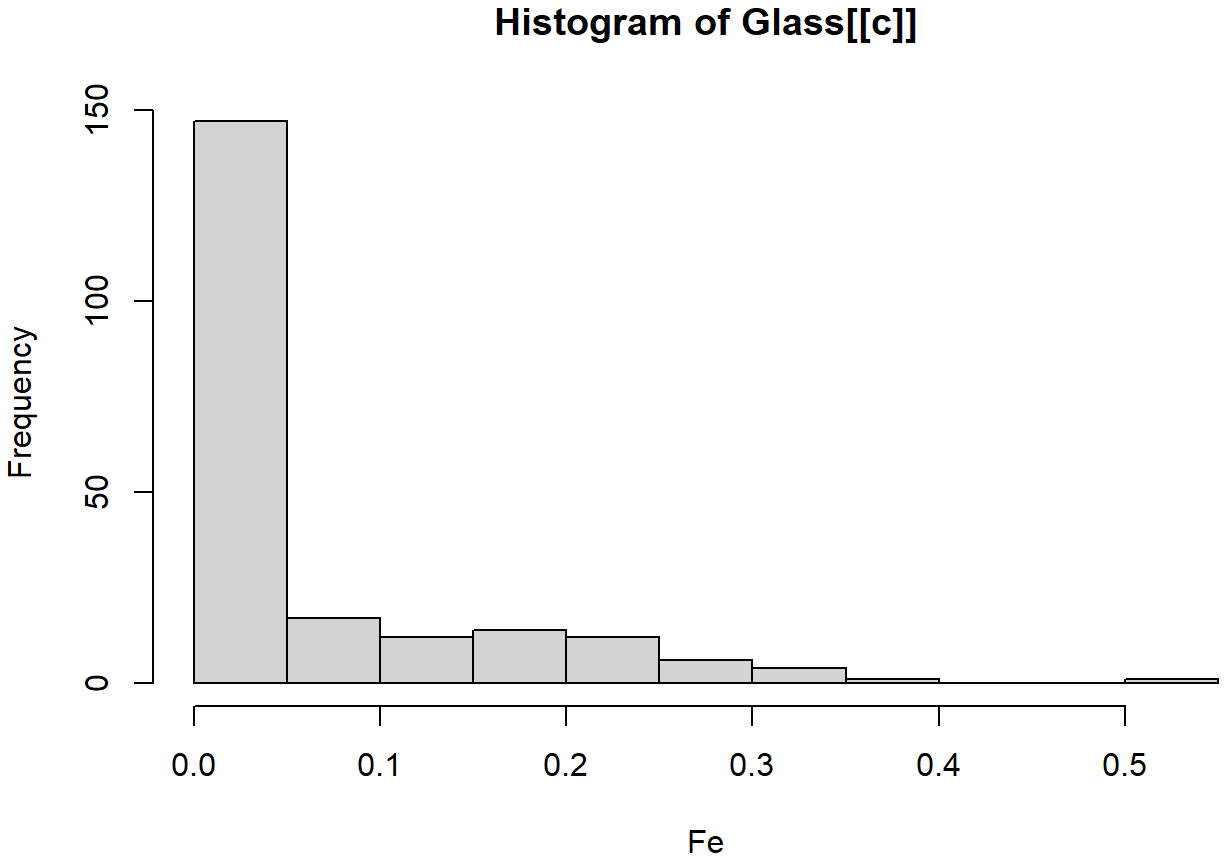

```
## [1] "Ca"
```


Histogram of Glass[[c]]
Frequency
0
20
40
60
80
100
120
6
8
10
12
14
16
Ca

```
## [1] "Ba"
```


Histogram of Glass[[c]]
Frequency
0
50
100
150
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Ba

```
## [1] "Fe"
```




```
## [1] "Type column is not numeric"
```

```
## SKEWNESS
print("RI")
```

```
## [1] "RI"
```

```
print(skewness(Glass$RI))
```

```
## [1] 1.614015
```

```
print("Na")
```

```
## [1] "Na"
```

```
print(skewness(Glass$Na))
```

```
## [1] 0.4509917
```

```
print("Mg")
```

```
## [1] "Mg"
```

```
print(skewness(Glass$Mg))
```

```
## [1] -1.144465
```

```
print("Al")
```

```
## [1] "Al"
```

```
print(skewness(Glass$Al))
```

```
## [1] 0.9009179
```

```
print("Si")
```

```
## [1] "Si"
```

```
print(skewness(Glass$Si))
```

```
## [1] -0.7253173
```

```
print("K")
```

```
## [1] "K"
```

```
print(skewness(Glass$K))
```

```
## [1] 6.505636
```

```
print("Ca")
```

```
## [1] "Ca"
```

```
print(skewness(Glass$Ca))
```

```
## [1] 2.032677
```

```
print("Ba")
```

```
## [1] "Ba"
```

```
print(skewness(Glass$Ba))
```

```
## [1] 3.392431
```

```
print("Fe")
```

```
## [1] "Fe"
```

```
print(skewness(Glass$Fe))
```

```
## [1] 1.742007
```

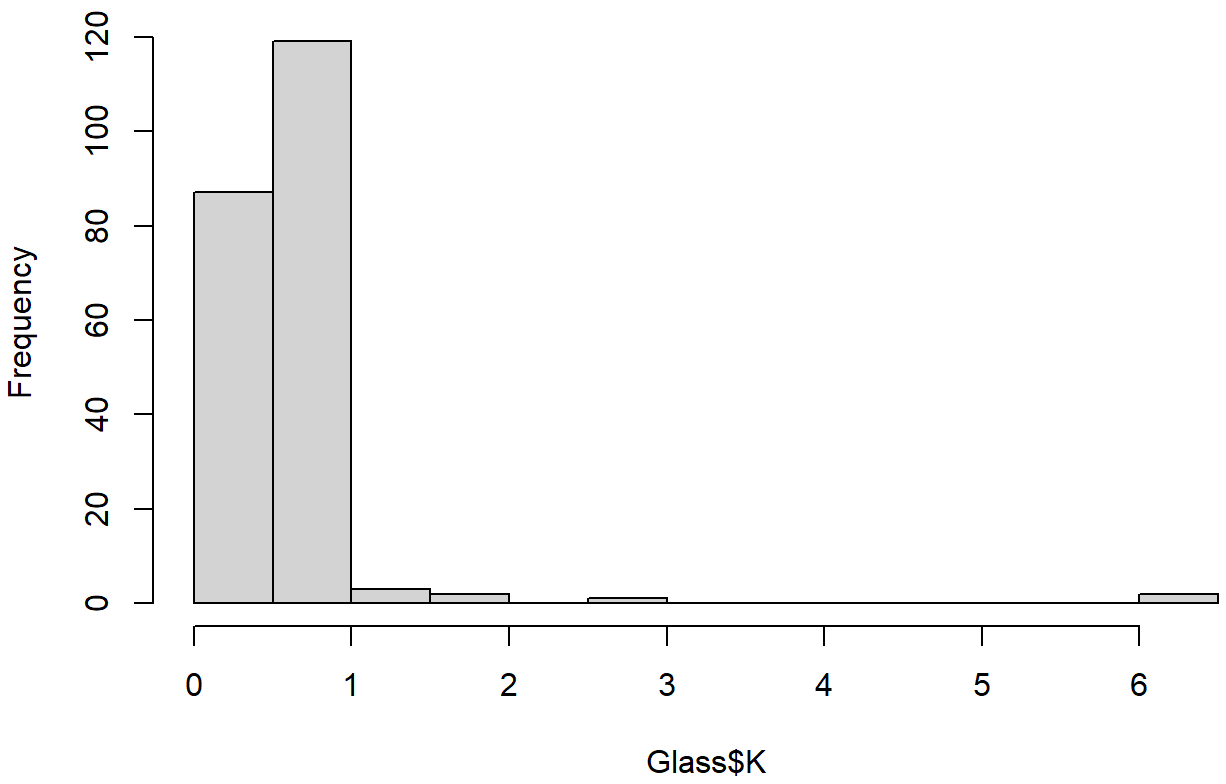
####(c) Are there any relevant transformations of one or more predictors that might improve the classification model? Yes, i used Box Cox on each of the predictors in order to transform and help their distrbutions. See below.

```
# Transformations for the skewed variables.
#install.packages("caret")
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.4.2
```

```
## Loading required package: lattice
```

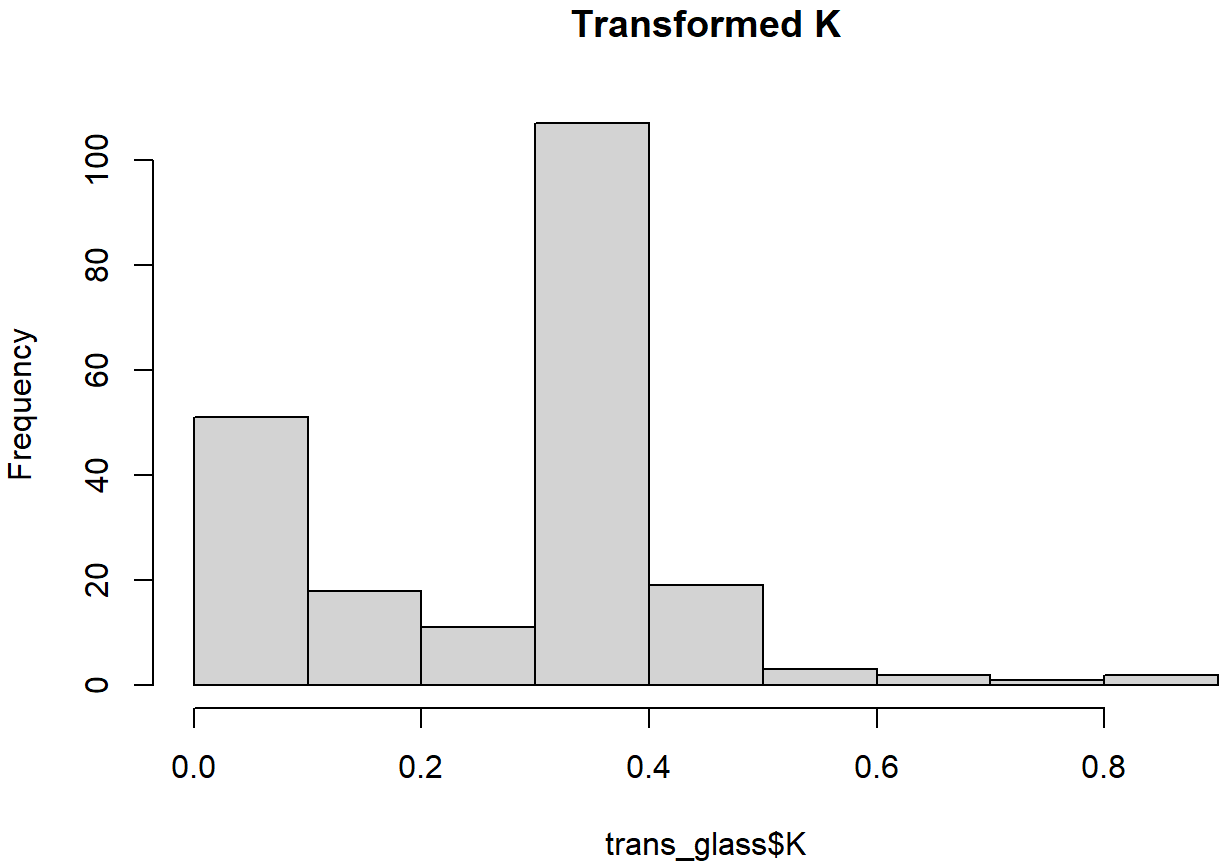
```
trans_glass <- Glass

## K
hist(Glass$K, main="Original K")
```



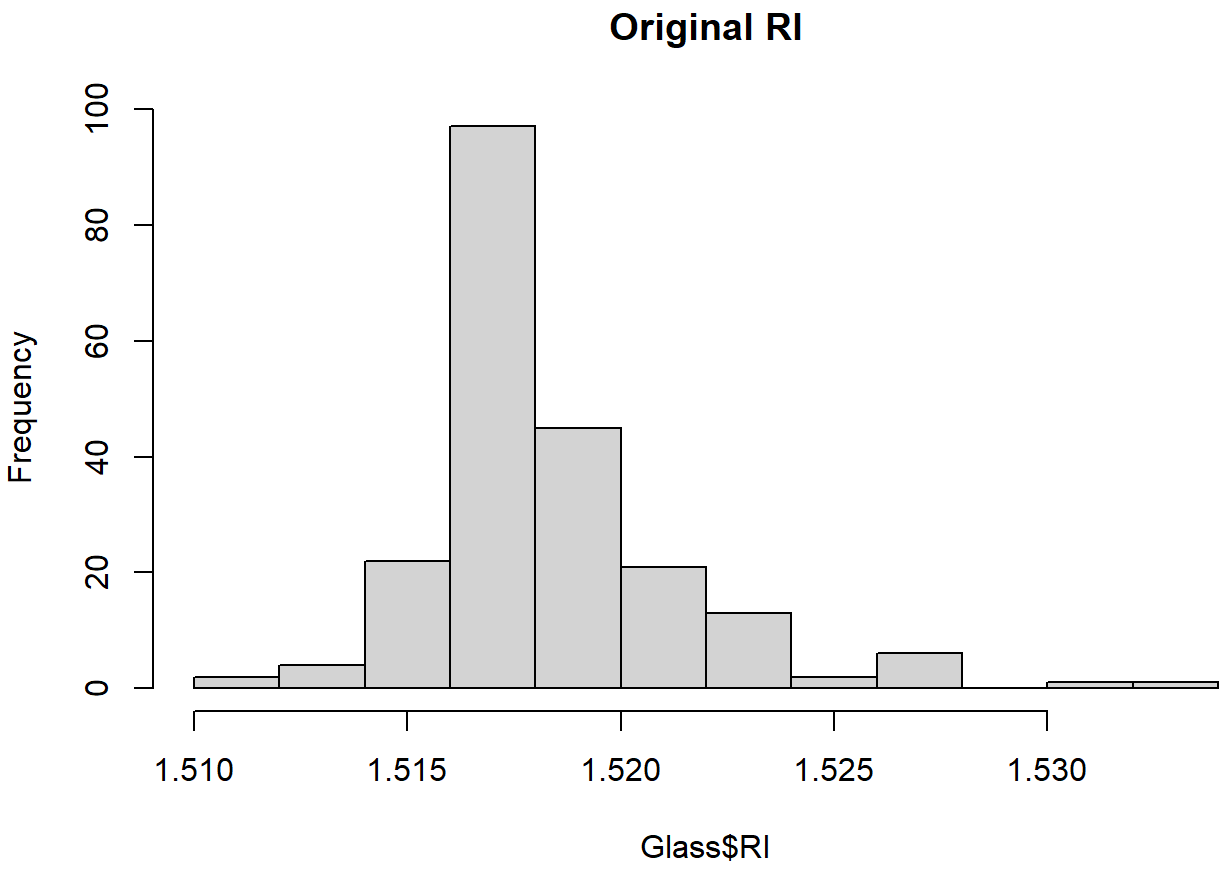

```
trans_k<- BoxCoxTrans(Glass$K+1) # Taking care of zeros
print(trans_k$lambda) ## Optimal Lambda is -1
```

```
## [1] -1
```

```
trans_glass$K <- ((trans_glass$K + 1)^trans_k$lambda - 1) / trans_k$lambda
hist(trans_glass$K, main="Transformed K")
```



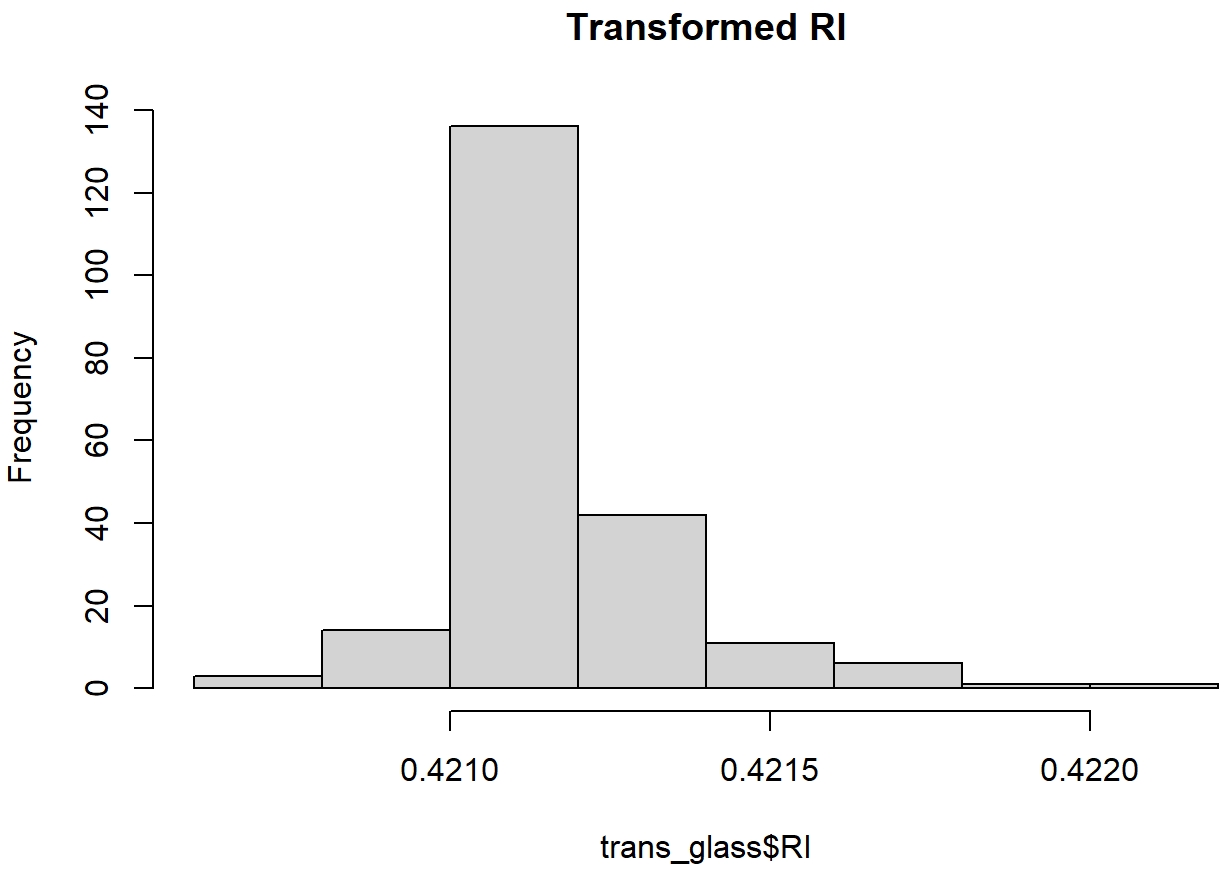

```
## RI
hist(Glass$RI, main="Original RI")
```



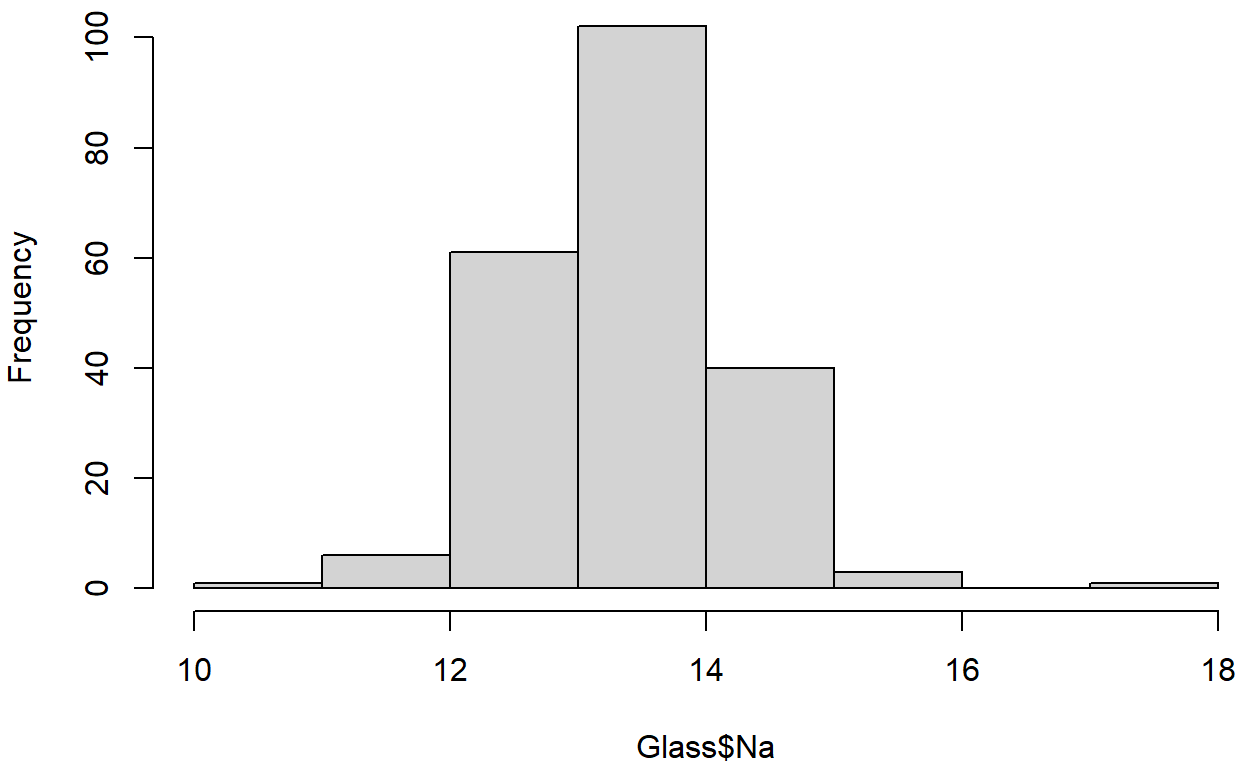

```
trans_RI<- BoxCoxTrans(Glass$RI+1) # Taking care of zeros
print(trans_RI$lambda) ## Optimal Lambda is -2
```

```
## [1] -2
```

```
trans_glass$RI <- ((trans_glass$RI + 1)^trans_RI$lambda - 1) / trans_RI$lambda
hist(trans_glass$RI, main="Transformed RI")
```



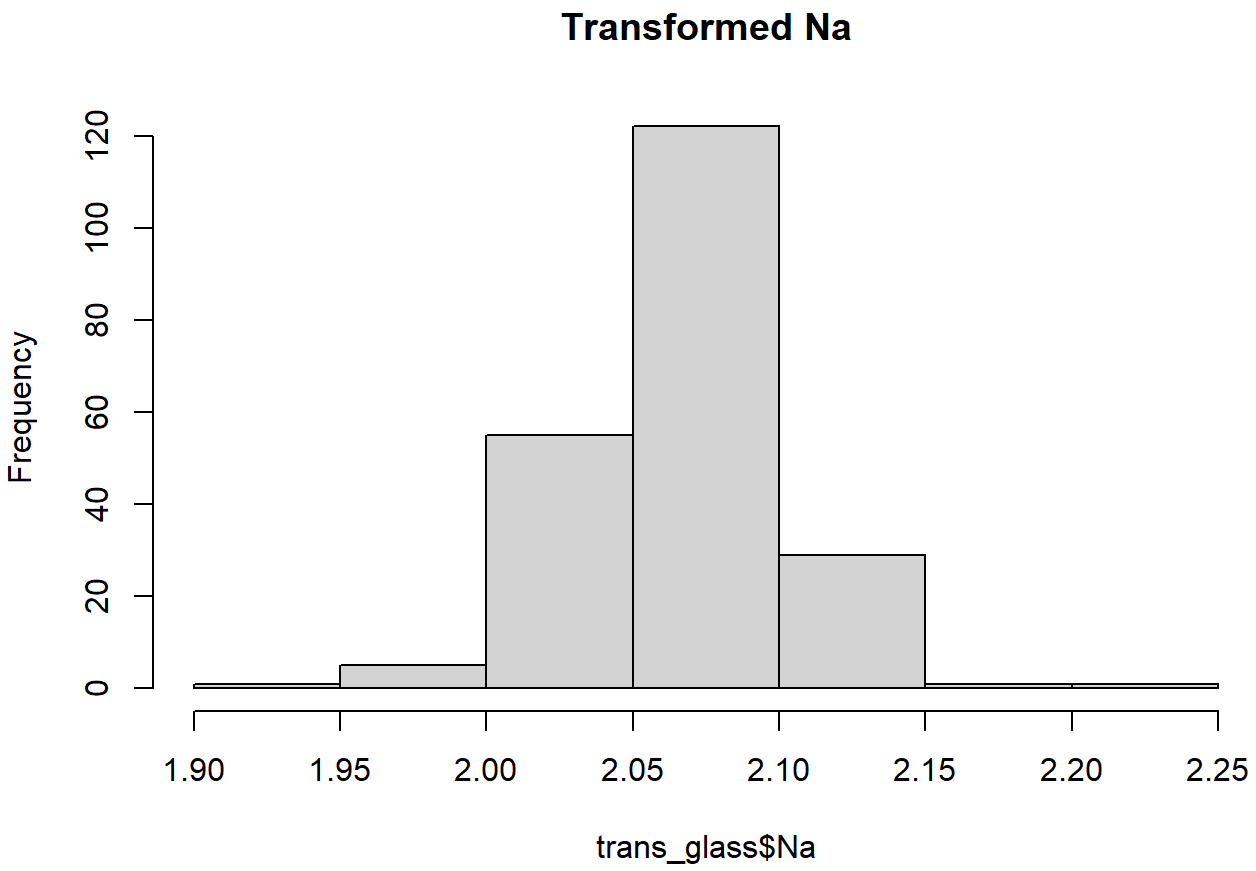

```
## Na
hist(Glass$Na, main="Original Na")
```



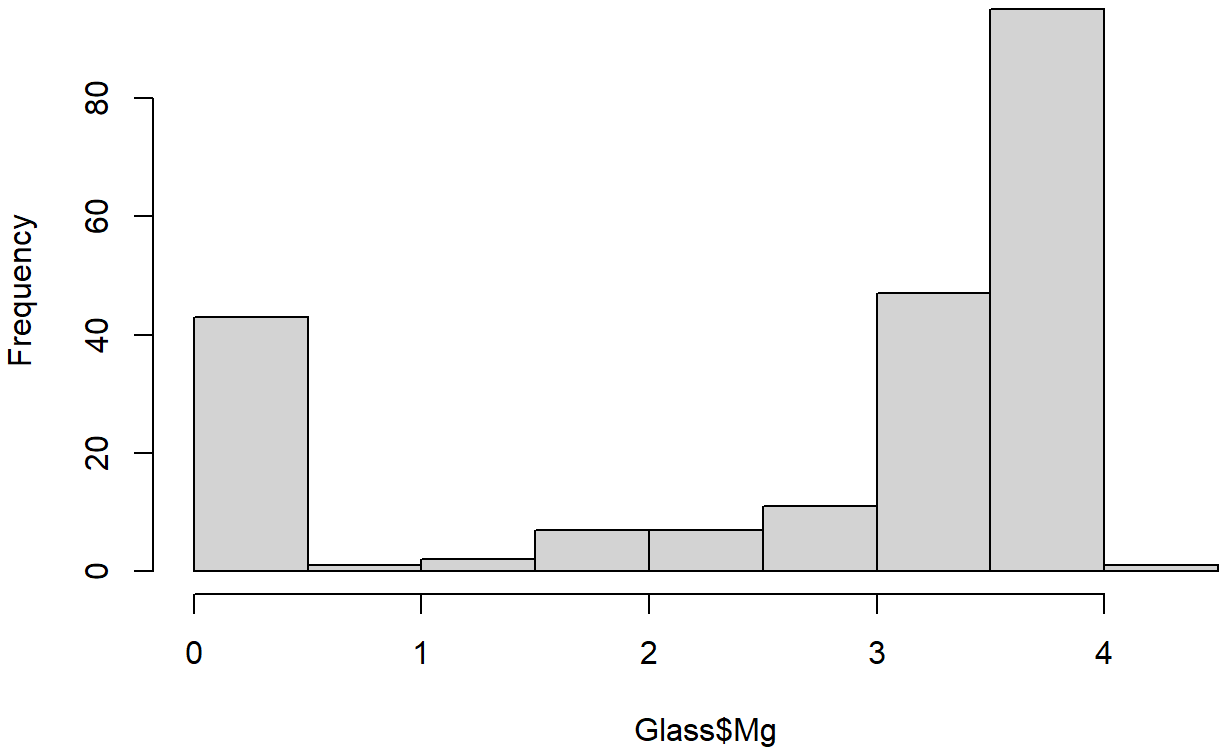

```
trans_Na<- BoxCoxTrans(Glass$Na+1) # Taking care of zeros
print(trans_Na$lambda) ## Optimal Lambda is -0.2
```

```
## [1] -0.2
```

```
trans_glass$Na <- ((trans_glass$Na + 1)^trans_Na$lambda - 1) / trans_Na$lambda
hist(trans_glass$Na, main="Transformed Na")
```



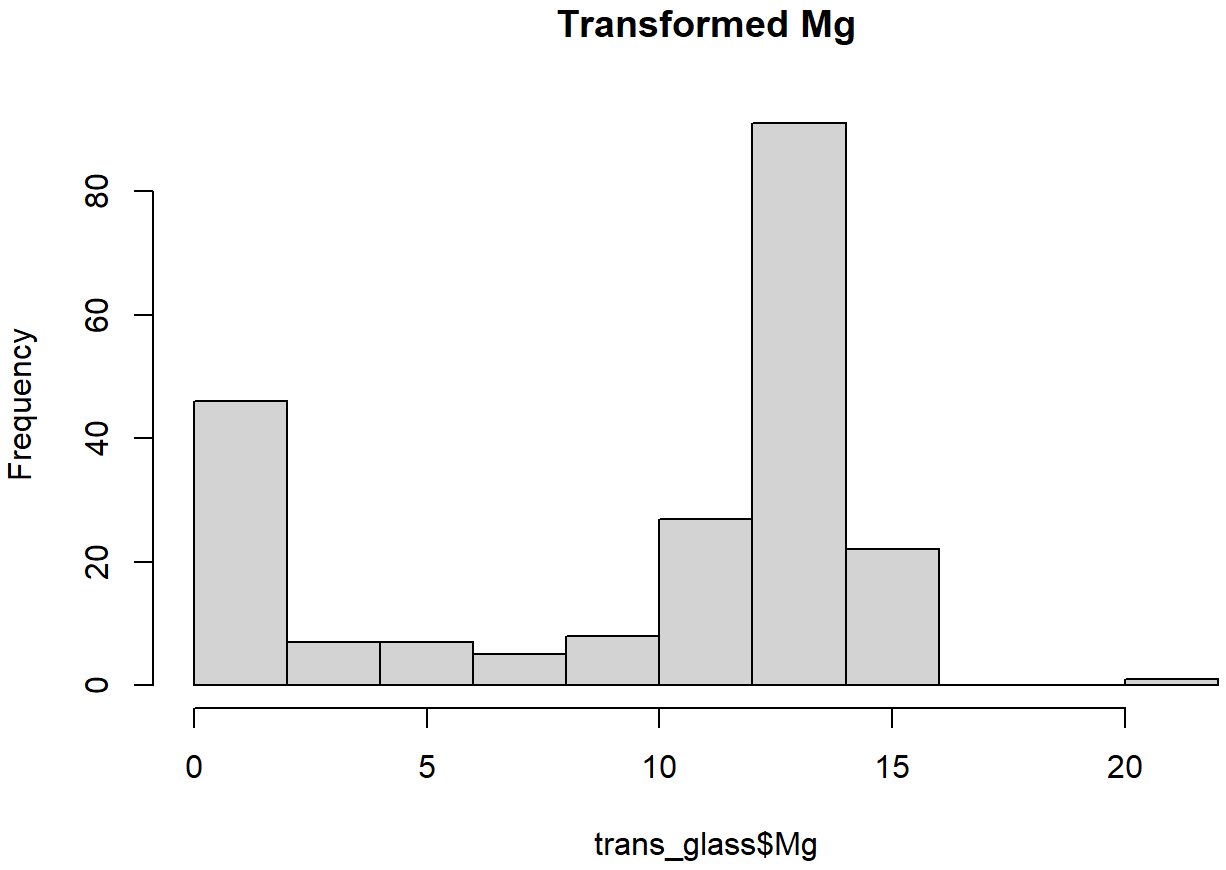

```
## Mg
hist(Glass$Mg, main="Original Mg")
```



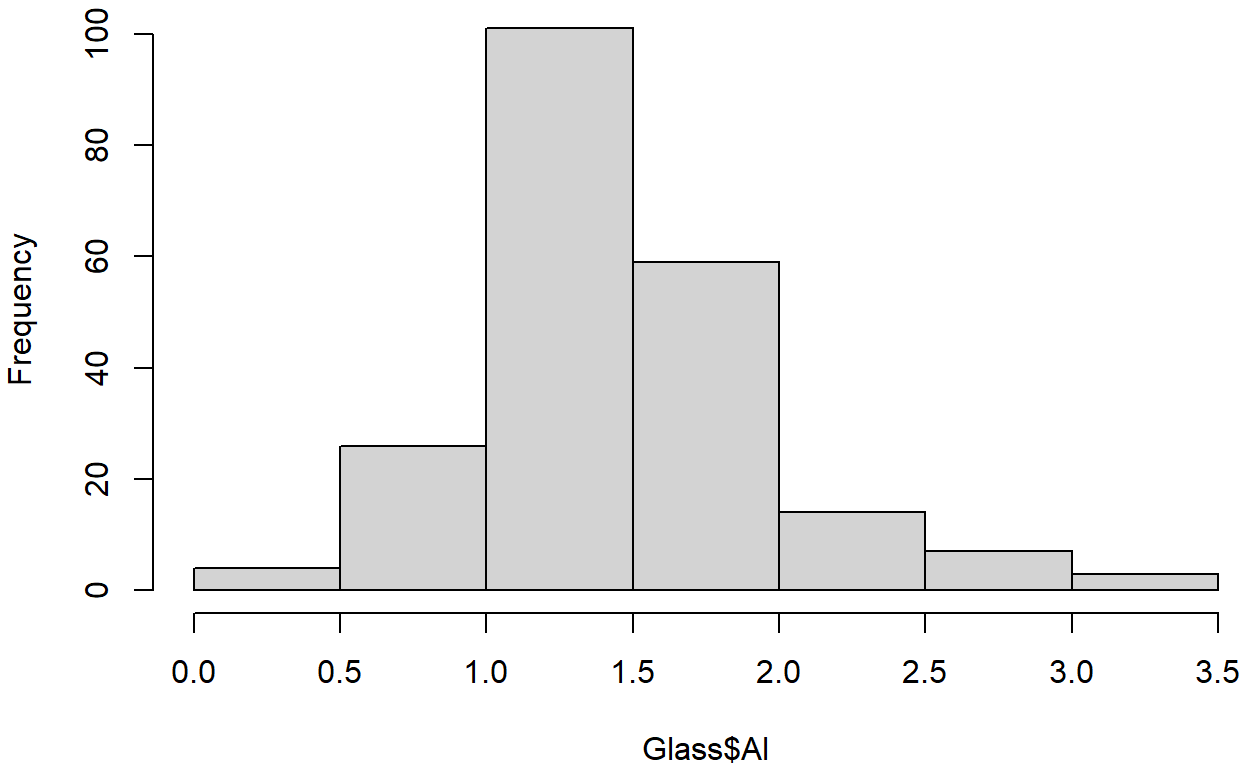

```
trans_Mg<- BoxCoxTrans(Glass$Mg+1) # Taking care of zeros
print(trans_Mg$lambda) ## Optimal Lambda is 2
```

```
## [1] 2
```

```
trans_glass$Mg <- (trans_glass$Mg)^2
hist(trans_glass$Mg, main="Transformed Mg")
```



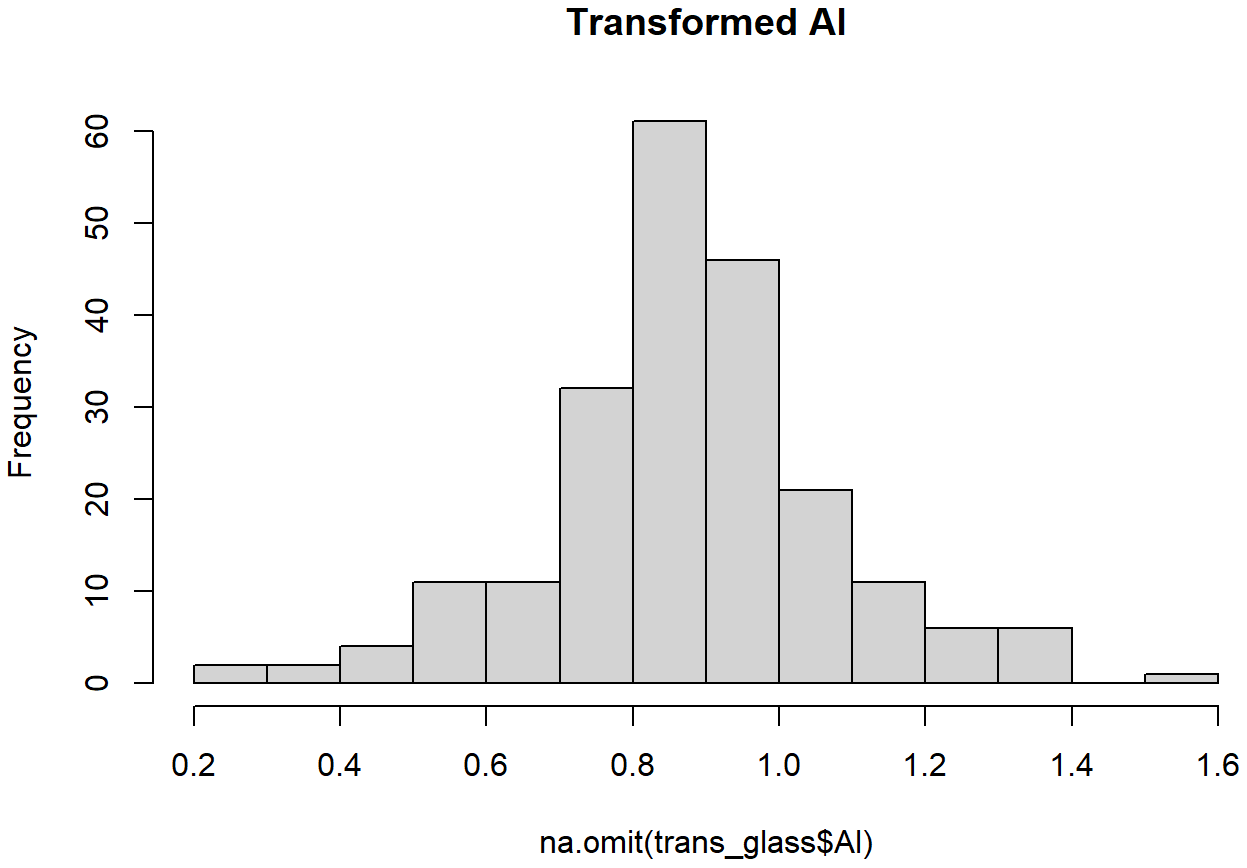

```
## Al
hist(Glass$Al, main="Original Al")
```



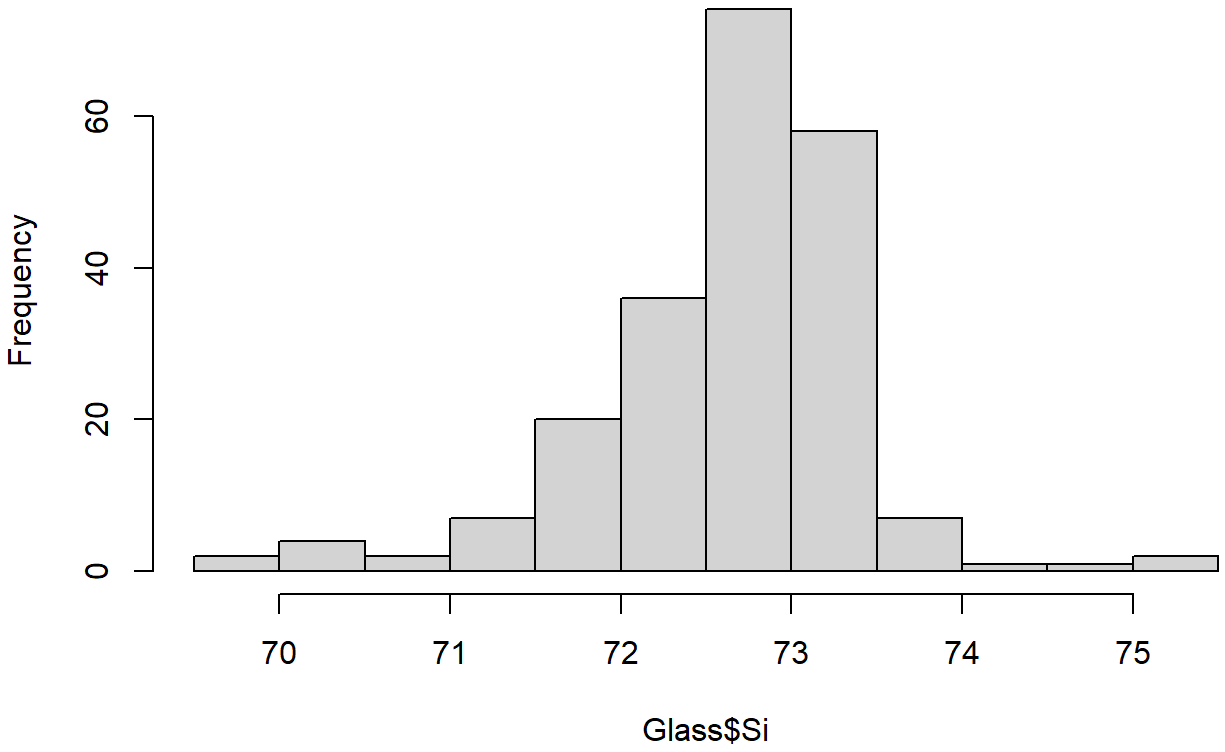

```
trans_Al<- BoxCoxTrans(Glass$Al+1) # Taking care of zeros
print(trans_Al$lambda) ## Optimal Lambda is
```

```
## [1] 0
```

```
trans_glass$Al <- log(trans_glass$Al + 1)
hist(na.omit(trans_glass$Al), main="Transformed Al")
```



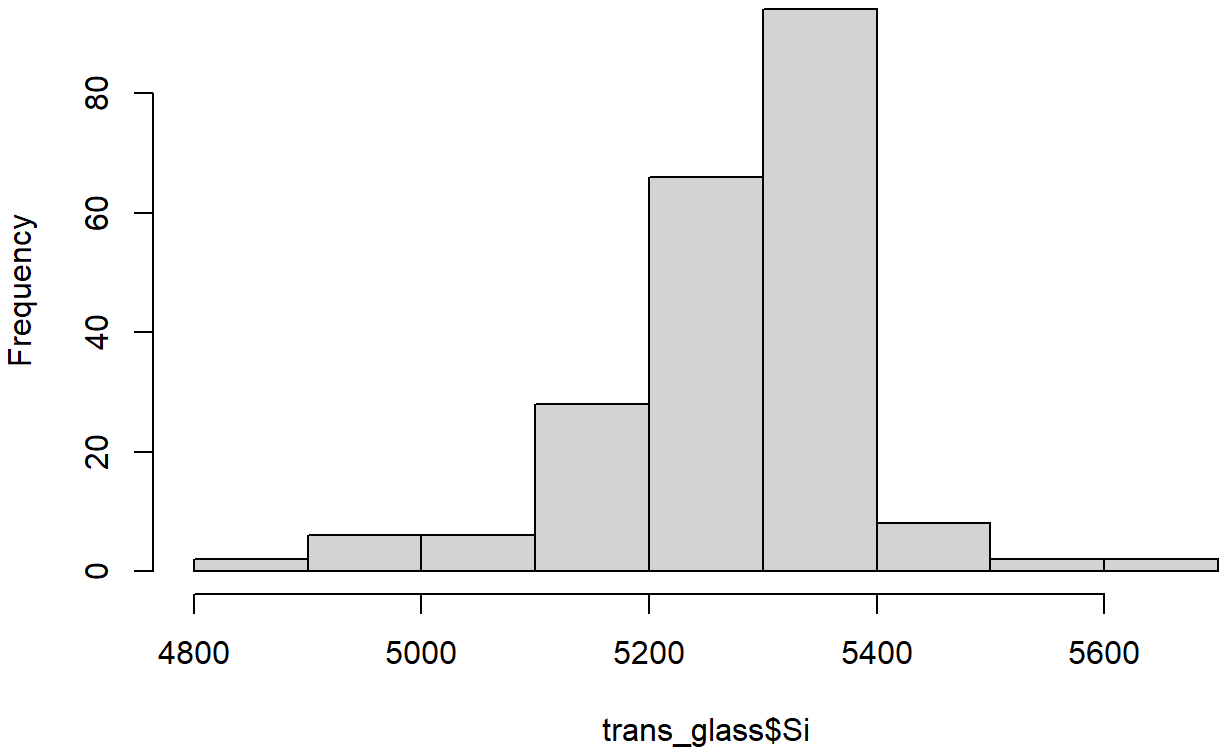

```
## Si
hist(Glass$Si, main="Original Si")
```



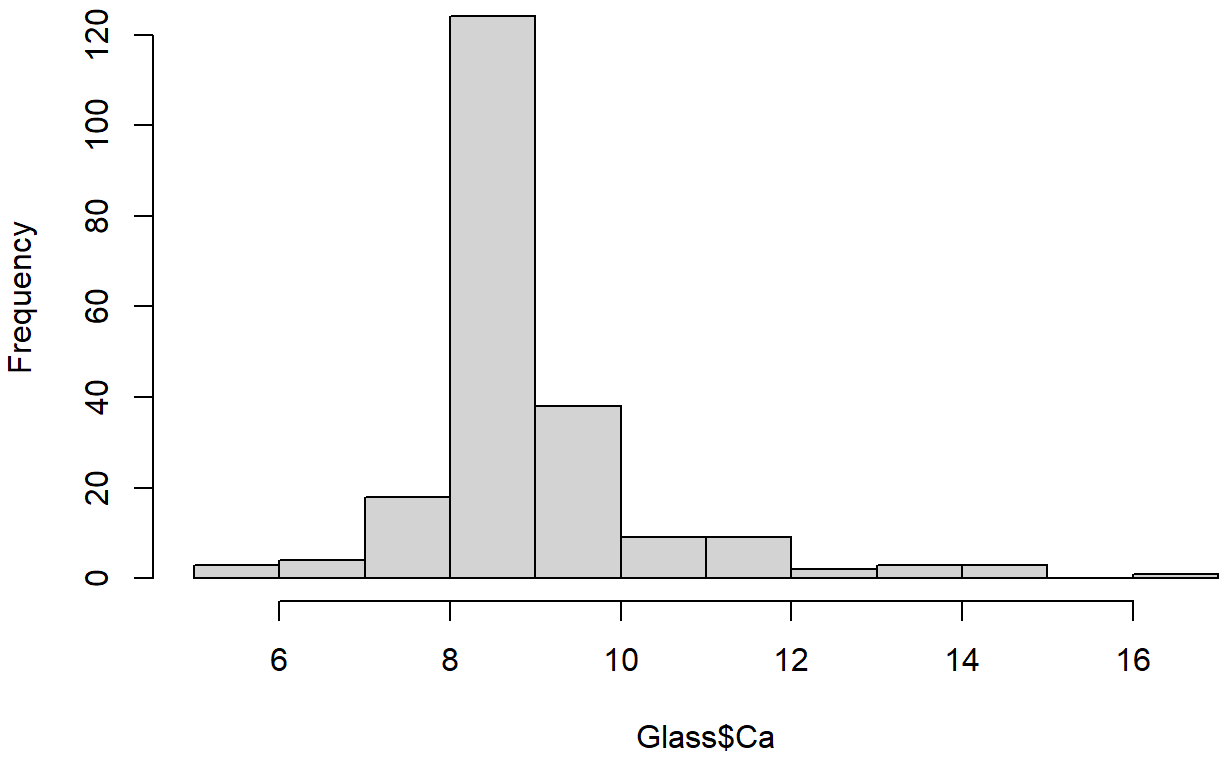

```
trans_Si <- BoxCoxTrans(Glass$Si + 1)
print(trans_Si$lambda)  ##2
```

```
## [1] 2
```

```
trans_glass$Si <- (trans_glass$Si)^2
hist(trans_glass$Si, main="Transformed Si")
```



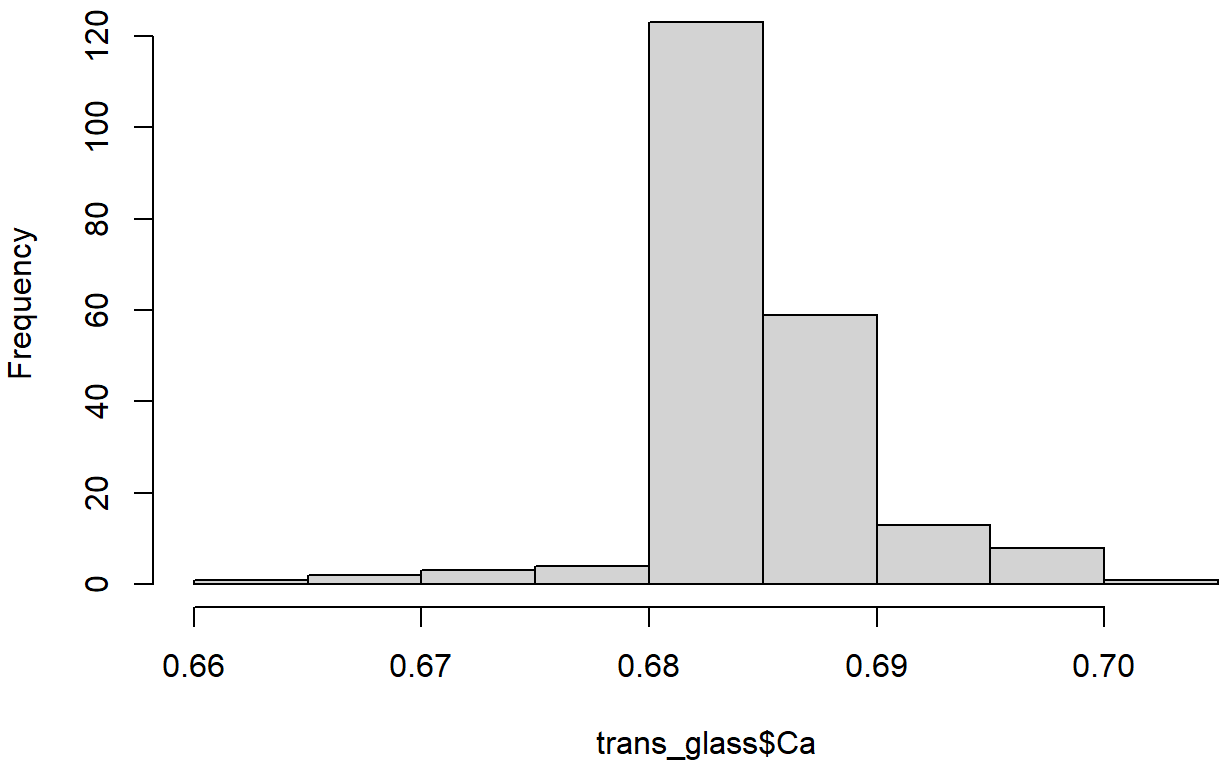

```
## Ca
hist(Glass$Ca, main="Original Ca")
```



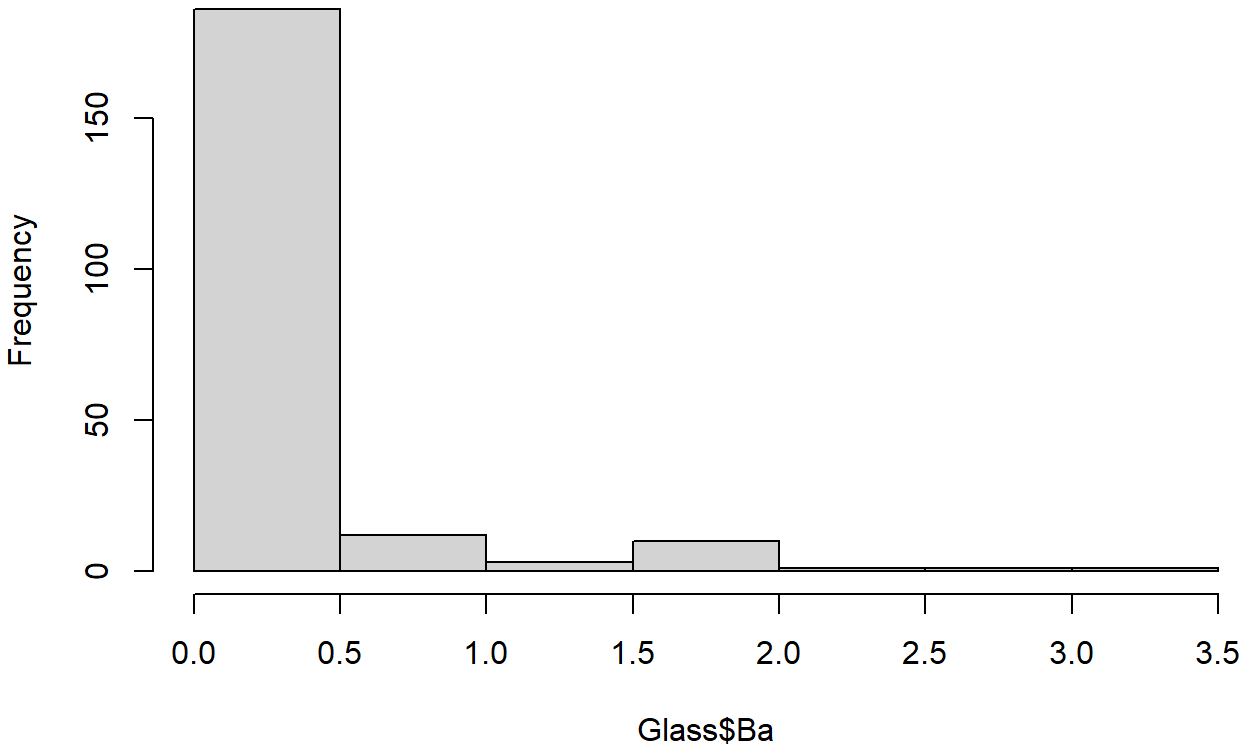

```
trans_Ca <- BoxCoxTrans(Glass$Ca + 1)
print(trans_Ca$lambda)
```

```
## [1] -1.4
```

```
trans_glass$Ca <- ((Glass$Ca + 1)^trans_Ca$lambda - 1) / trans_Ca$lambda
hist(trans_glass$Ca, main="Transformed Ca")
```



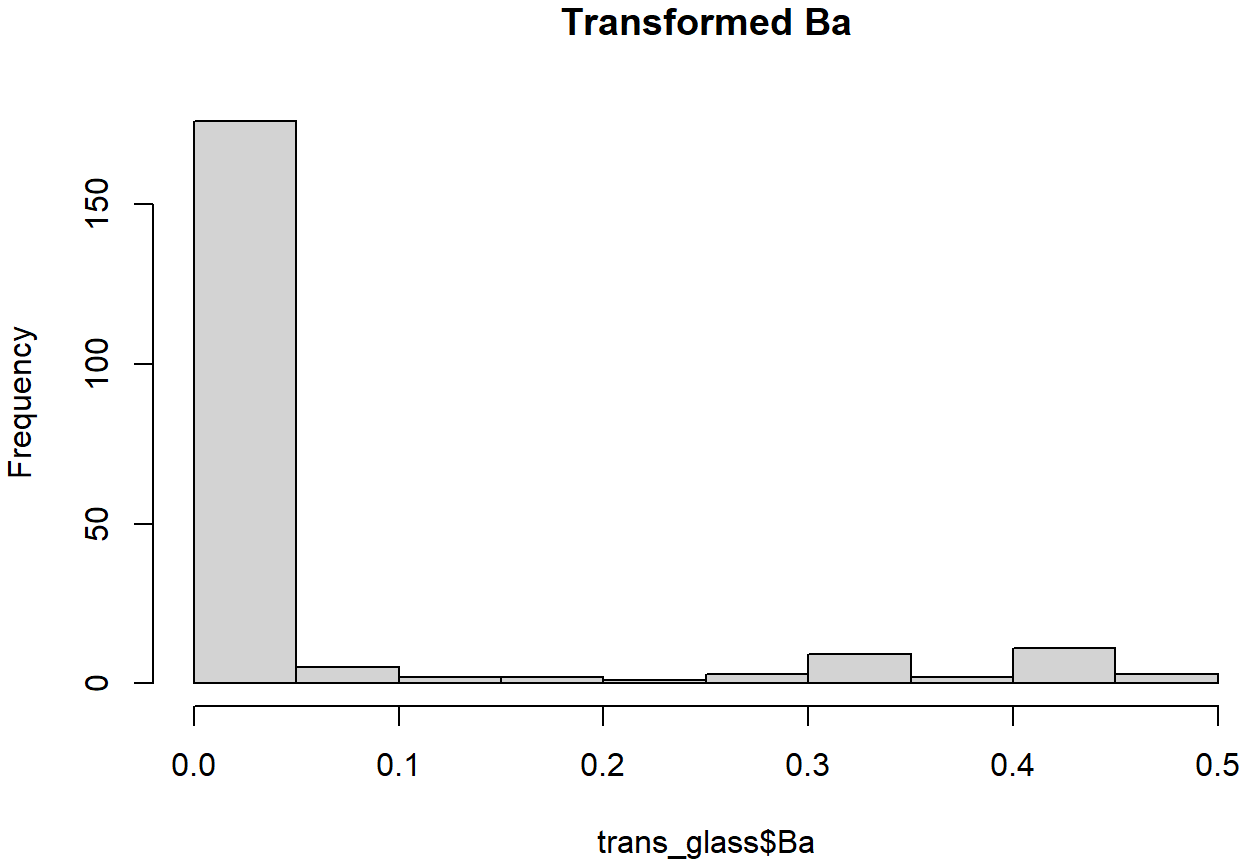

```
## Ba
hist(Glass$Ba, main="Original Ba")
```



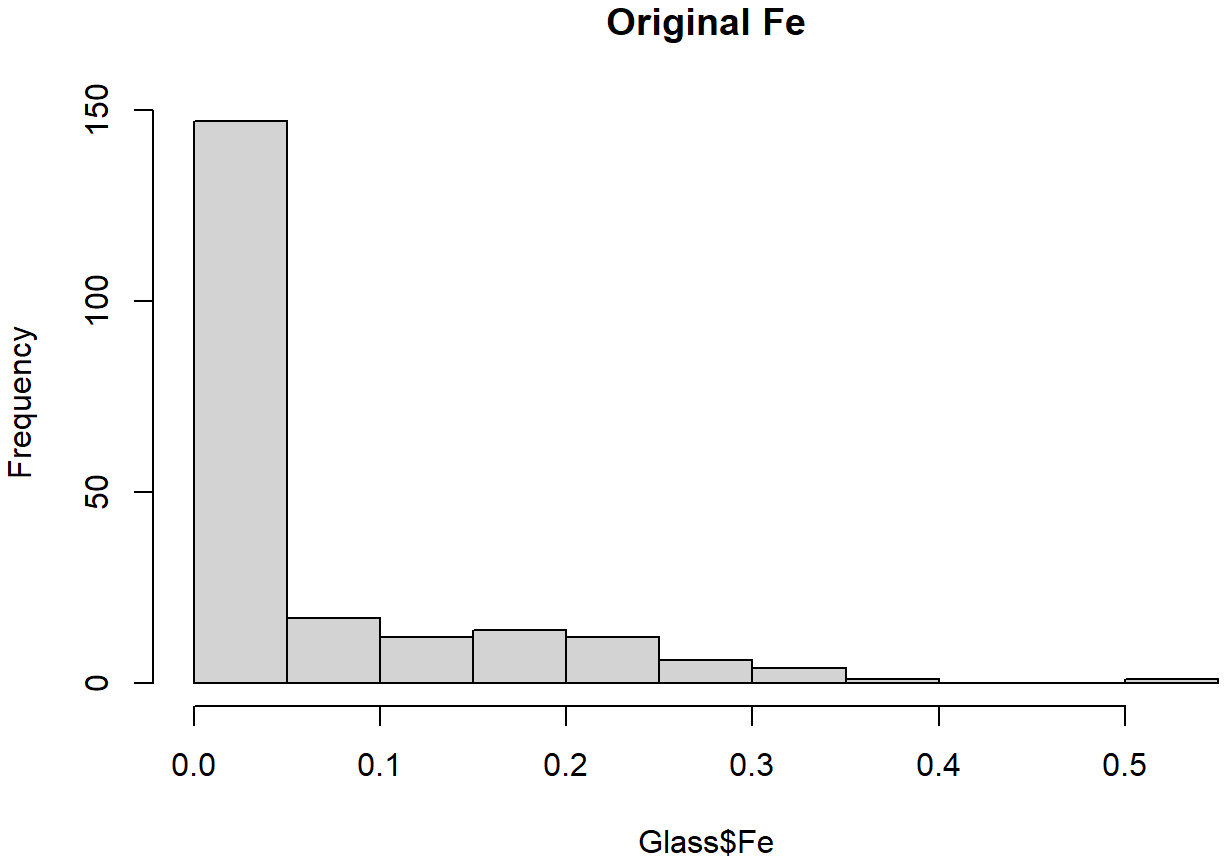

```
trans_Ba <- BoxCoxTrans(Glass$Ba + 1)
print(trans_Ba$lambda)
```

```
## [1] -2
```

```
trans_glass$Ba <- ((Glass$Ba + 1)^trans_Ba$lambda - 1) / trans_Ba$lambda
hist(trans_glass$Ba, main="Transformed Ba")
```



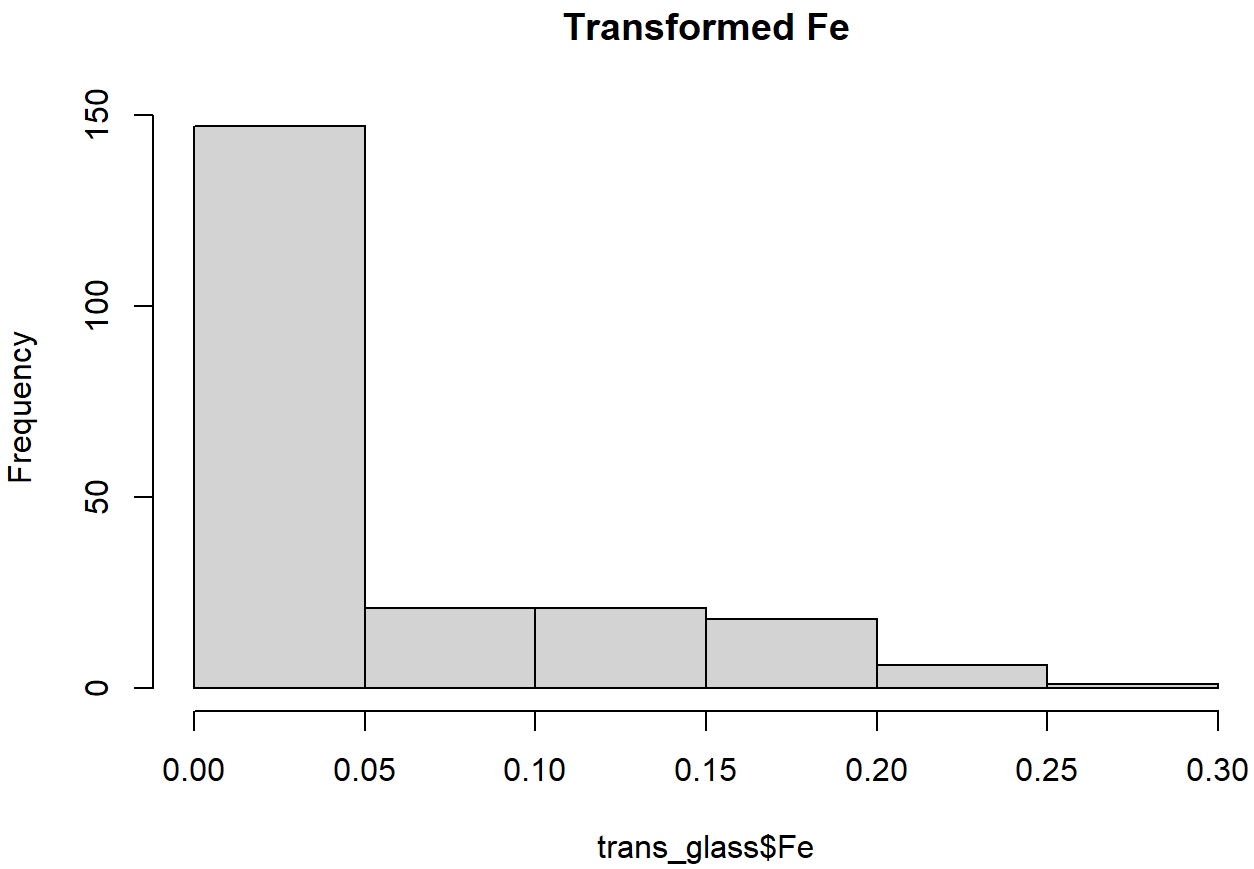

```
## Fe
hist(Glass$Fe, main="Original Fe")
```




```
trans_Fe <- BoxCoxTrans(Glass$Fe + 1)
print(trans_Fe$lambda)
```

```
## [1] -2
```

```
trans_glass$Fe <- ((Glass$Fe + 1)^trans_Fe$lambda - 1) / trans_Fe$lambda
hist(trans_glass$Fe, main="Transformed Fe")
```



3.2.

The soybean data can also be found at the UC Irvine Machine Learning Repository. Data were collected to predict disease in 683 soybeans. The 35 predictors are mostly categorical and include information on the environmental conditions (e.g., temperature, precipitation) and plant conditions (e.g., left spots, mold growth). The outcome labels consist of 19 distinct classes.The data can be loaded via:

```
#library(mlbench)
data(Soybean)
## See ?Soybean for details
#?Soybean
```

(a) Investigate the frequency distributions for the categorical predictors. Are any of the distributions degenerate in the ways discussed earlier in this chapter?

Yes, many of the columns in this table do not have a large amount of variability, the data is numeric categorical dummy variables. Therefore it's a hurdle when attempting to model any type of variability.

```
#head(Soybean)
for (c in colnames(Soybean)){
  print(c)
 print(table(Soybean[[c]]))
}
```

```
## [1] "Class"
##
##                 2-4-d-injury         alternarialeaf-spot
##                           16                          91
##                  anthracnose            bacterial-blight
##                           44                          20
##           bacterial-pustule                  brown-spot
##                           20                          92
##              brown-stem-rot               charcoal-rot
##                           44                          20
##               cyst-nematode diaporthe-pod-&-stem-blight
##                           14                          15
##       diaporthe-stem-canker               downy-mildew
##                           20                          20
##          frog-eye-leaf-spot            herbicide-injury
##                           91                           8
##      phyllosticta-leaf-spot            phytophthora-rot
##                           20                          88
##              powdery-mildew           purple-seed-stain
##                           20                          20
##        rhizoctonia-root-rot
##                           20
## [1] "date"
##
##   0   1   2   3   4   5   6
##  26  75  93 118 131 149  90
## [1] "plant.stand"
##
##   0   1
## 354 293
## [1] "precip"
##
##   0   1   2
##  74 112 459
## [1] "temp"
##
##   0   1   2
##  80 374 199
## [1] "hail"
##
##   0   1
## 435 127
## [1] "crop.hist"
##
##   0   1   2   3
##  65 165 219 218
## [1] "area.dam"
##
##   0   1   2   3
## 123 227 145 187
## [1] "sever"
```

```
##
##   0   1   2
## 195 322  45
## [1] "seed.tmt"
##
##   0   1   2
## 305 222  35
## [1] "germ"
##
##   0   1   2
## 165 213 193
## [1] "plant.growth"
##
##   0   1
## 441 226
## [1] "leaves"
##
##   0   1
##  77 606
## [1] "leaf.halo"
##
##   0   1   2
## 221  36 342
## [1] "leaf.marg"
##
##   0   1   2
## 357  21 221
## [1] "leaf.size"
##
##   0   1   2
##  51 327 221
## [1] "leaf.shread"
##
##   0   1
## 487  96
## [1] "leaf.malf"
##
##   0   1
## 554  45
## [1] "leaf.mild"
##
##   0   1   2
## 535  20  20
## [1] "stem"
##
##   0   1
## 296 371
## [1] "lodging"
##
##   0   1
## 520  42
```

```
## [1] "stem.cankers"
##
##   0   1   2   3
## 379  39  36 191
## [1] "canker.lesion"
##
##   0   1   2   3
## 320  83 177  65
## [1] "fruiting.bodies"
##
##   0   1
## 473 104
## [1] "ext.decay"
##
##   0   1   2
## 497 135  13
## [1] "mycelium"
##
##   0   1
## 639   6
## [1] "int.discolor"
##
##   0   1   2
## 581  44  20
## [1] "sclerotia"
##
##   0   1
## 625  20
## [1] "fruit.pods"
##
##   0   1   2   3
## 407 130  14  48
## [1] "fruit.spots"
##
##   0   1   2   4
## 345  75  57 100
## [1] "seed"
##
##   0   1
## 476 115
## [1] "mold.growth"
##
##   0   1
## 524  67
## [1] "seed.discolor"
##
##   0   1
## 513  64
## [1] "seed.size"
##
##   0   1
```

```
## 532  59
## [1] "shriveling"
##
##   0   1
## 539  38
## [1] "roots"
##
##   0   1   2
## 551  86  15
```

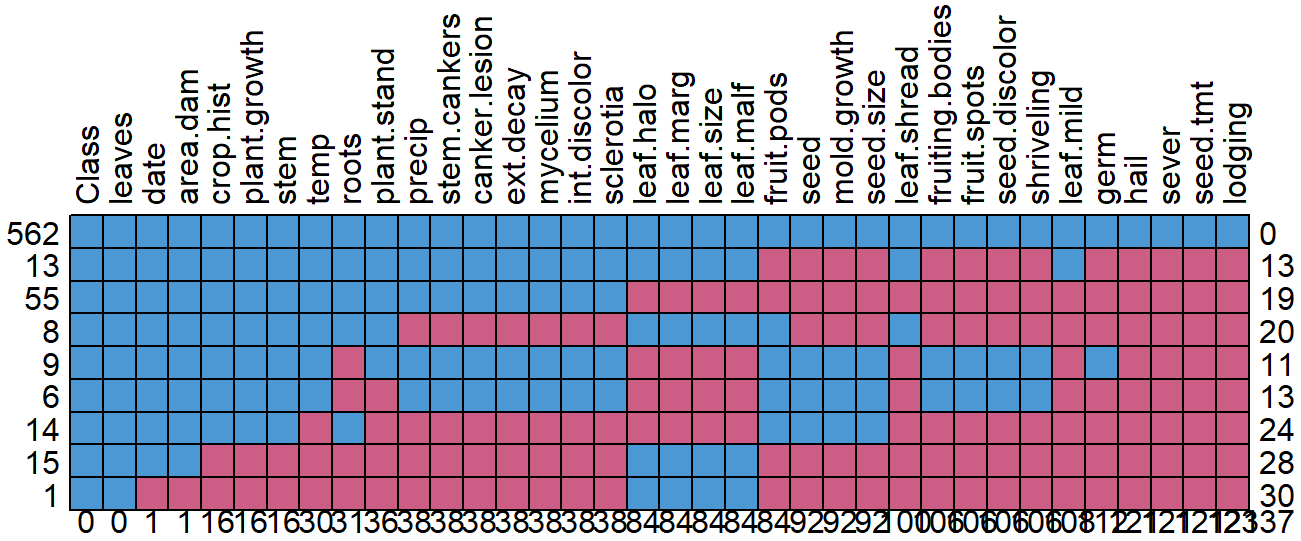
```
#The following columnd only have binary values, that is there is only 1 or 0 values present i
n the column:
#  plant.stand
#  hail
#  plant.growth
#  leaves
#  leaf.shread
#  leaf.malf
#  stem
#  lodging
#  fruiting.bodies
#  mycellium
#  sclerotia
#  seed
#  seed.discolor
#  seed.size
#  shriveling

#These binary columns also have a disproportionate amount of zeros when compared to 1's, exce
pt for plant.stand, stem,  plant.growth.
# plant.leaf has many more 1 values than 0.
```

(b) Roughly 18 % of the data are missing. Are there particular predictors that are more likely to be missing? Is the pattern of missing data related to the classes?

The predictor columns that are most likely to be null are: hail, sever, seet.tmt, lodging. Those with no nulls are:Class, Leaves Those with nulls, but have the fewest amount: date,area.dam, crop.hist,stem,plant.growth.

When looking at the Class the nulls are mostly limited to a handfull of class values, so these should probably be removed. The class values that have all of the nulls are: phytophthora-rot, 2-4-d-injury,cyst-nematode,diaporthe-pod-&-stem-blight, and herbicide-injury.

```
## Using the MICE package to display the null coverage.
md.pattern(Soybean,rotate.names = TRUE)
```


Class
leaves
date
area.dam
crop.hist
plant.growth
stem
temp
roots
plant.stand
precip
stem.cankers
canker.lesion
ext.decay
mycelium
int.discolor
sclerotia
leaf.halo
leaf.marg
leaf.size
leaf.malf
fruit.pods
seed
mold.growth
seed.size
leaf.shread
fruiting.bodies
fruit.spots
seed.discolor
shriveling
leaf.mild
germ
hail
sever
seed.tmt
lodging
562
13
55
8
9
6
14
15
1
0
13
19
20
11
13
24
28
30
0 0 1 1 16 16 16 30 31 36 38 38 38 38 38 38 38 38 48 48 48 48 48 49 29 29 21 00 00 60 60 60 60 60 81 12 21 21 21 21 23 37

```
##     Class leaves date area.dam crop.hist plant.growth stem temp roots
## 562     1      1    1        1         1            1    1    1     1
## 13      1      1    1        1         1            1    1    1     1
## 55      1      1    1        1         1            1    1    1     1
## 8       1      1    1        1         1            1    1    1     1
## 9       1      1    1        1         1            1    1    1     0
## 6       1      1    1        1         1            1    1    1     0
## 14      1      1    1        1         1            1    1    0     1
## 15      1      1    1        1         0            0    0    0     0
## 1       1      1    0        0         0            0    0    0     0
##         0      0    1        1        16           16   16   30    31
##     plant.stand precip stem.cankers canker.lesion ext.decay mycelium
## 562           1      1            1             1         1        1
## 13            1      1            1             1         1        1
## 55            1      1            1             1         1        1
## 8             1      0            0             0         0        0
## 9             1      1            1             1         1        1
## 6             0      1            1             1         1        1
## 14            0      0            0             0         0        0
## 15            0      0            0             0         0        0
## 1             0      0            0             0         0        0
##              36     38           38            38        38       38
##     int.discolor sclerotia leaf.halo leaf.marg leaf.size leaf.malf fruit.pods
## 562            1         1         1         1         1         1          1
## 13             1         1         1         1         1         1          0
## 55             1         1         0         0         0         0          0
## 8              0         0         1         1         1         1          1
## 9              1         1         0         0         0         0          1
## 6              1         1         0         0         0         0          1
## 14             0         0         0         0         0         0          1
## 15             0         0         1         1         1         1          0
## 1              0         0         1         1         1         1          0
##               38        38        84        84        84        84         84
##     seed mold.growth seed.size leaf.shread fruiting.bodies fruit.spots
## 562    1           1         1           1               1          1
## 13     0           0         0           1               0          0
## 55     0           0         0           0               0          0
## 8      0           0         0           1               0          0
## 9      1           1         1           0               1          1
## 6      1           1         1           0               1          1
## 14     1           1         1           0               0          0
## 15     0           0         0           0               0          0
## 1      0           0         0           0               0          0
##       92          92        92         100             106        106
##     seed.discolor shriveling leaf.mild germ hail sever seed.tmt lodging
## 562             1          1         1    1    1     1        1       1    0
## 13              0          0         1    0    0     0        0       0   13
## 55              0          0         0    0    0     0        0       0   19
## 8               0          0         0    0    0     0        0       0   20
## 9               1          1         0    1    0     0        0       0   11
## 6               1          1         0    0    0     0        0       0   13
```

```
## 14              0          0         0    0    0     0        0       0   24
## 15              0          0         0    0    0     0        0       0   28
## 1               0          0         0    0    0     0        0       0   30
##               106        106       108  112  121   121      121     121 2337
```

```
## Most nulls: hail sever seet.tmt lodging

## Least NUlls: date,area.dam, crop.hist,stem,plant.growth

## NO Nulls : Class, Leaves
missing_counts <- Soybean |>
  group_by(Class) |>
  summarise(across(everything(), ~sum(is.na(.)), .names = "missing_{col}"))

missing_counts <- missing_counts %>%
  mutate(total_nulls = rowSums(across(starts_with("missing_"))))

print(nrow(Soybean)) #683
```

```
## [1] 683
```

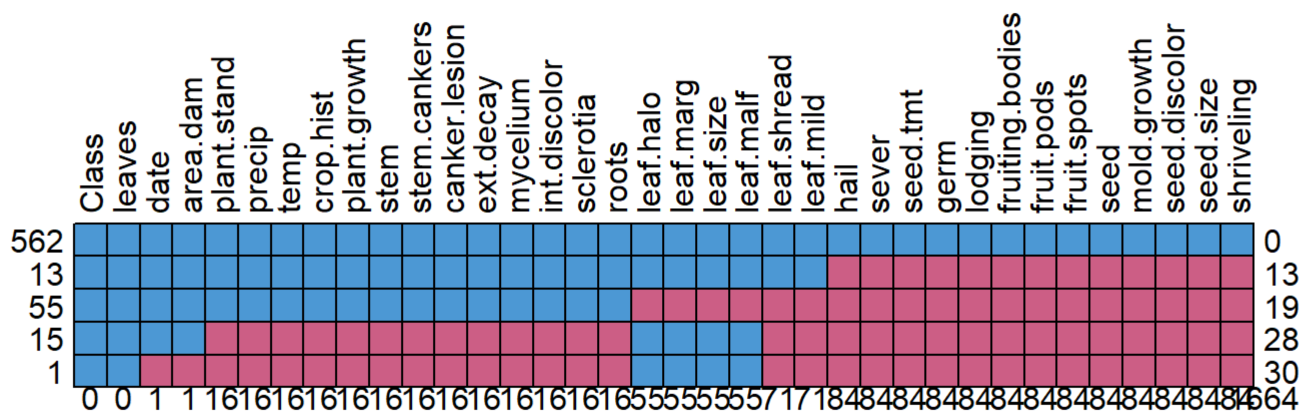
```
print(missing_counts |> select(Class, total_nulls) |>arrange(desc(total_nulls) ))
```

```
## # A tibble: 19 × 2
##    Class                       total_nulls
##    <fct>                             <dbl>
##  1 phytophthora-rot                   1214
##  2 2-4-d-injury                        450
##  3 cyst-nematode                       336
##  4 diaporthe-pod-&-stem-blight         177
##  5 herbicide-injury                    160
##  6 alternarialeaf-spot                   0
##  7 anthracnose                           0
##  8 bacterial-blight                      0
##  9 bacterial-pustule                     0
## 10 brown-spot                            0
## 11 brown-stem-rot                        0
## 12 charcoal-rot                          0
## 13 diaporthe-stem-canker                 0
## 14 downy-mildew                          0
## 15 frog-eye-leaf-spot                    0
## 16 phyllosticta-leaf-spot                0
## 17 powdery-mildew                        0
## 18 purple-seed-stain                     0
## 19 rhizoctonia-root-rot                  0
```

(c) Develop a strategy for handling missing data, either by eliminating predictors or imputation.

After removing the class values that are associated with the largest amount of null values in the data, the remaining class values has much fewer nulls. The max number of nulls in a column is 84, which is about 13% of the rows in the df. This level of nulls can then be imputed using what ever appropriate means to derive the values. However, the rows that are imputed shoudl be flagged with a second column to indicate where the data was imputed. For instance if the plant.growth column is imputed, the null value rows should be flagged with a "plat.growth_impute_flag" column, so that the analyst can keep track of where imputation was used.

```
no_nulls <- Soybean |> filter(!Class %in% c("phytophthora-rot, 2-4-d-injury","cyst-nematod
e","diaporthe-pod-&-stem-blight","herbicide-injury"))
md.pattern(no_nulls,rotate.names = TRUE)
```



```
##     Class leaves date area.dam plant.stand precip temp crop.hist plant.growth
## 562     1      1    1        1           1      1    1         1            1
## 13      1      1    1        1           1      1    1         1            1
## 55      1      1    1        1           1      1    1         1            1
## 15      1      1    1        1           0      0    0         0            0
## 1       1      1    0        0           0      0    0         0            0
##         0      0    1        1          16     16   16        16           16
##     stem stem.cankers canker.lesion ext.decay mycelium int.discolor sclerotia
## 562    1            1             1         1        1            1         1
## 13     1            1             1         1        1            1         1
## 55     1            1             1         1        1            1         1
## 15     0            0             0         0        0            0         0
## 1      0            0             0         0        0            0         0
##       16           16            16        16       16           16        16
##     roots leaf.halo leaf.marg leaf.size leaf.malf leaf.shread leaf.mild hail
## 562     1         1         1         1         1           1         1    1
## 13      1         1         1         1         1           1         1    0
## 55      1         0         0         0         0           0         0    0
## 15      0         1         1         1         1           0         0    0
## 1       0         1         1         1         1           0         0    0
##        16        55        55        55        55          71        71   84
##     sever seed.tmt germ lodging fruiting.bodies fruit.pods fruit.spots seed
## 562     1        1    1       1               1          1           1    1
## 13      0        0    0       0               0          0           0    0
## 55      0        0    0       0               0          0           0    0
## 15      0        0    0       0               0          0           0    0
## 1       0        0    0       0               0          0           0    0
##        84       84   84      84              84         84          84   84
##     mold.growth seed.discolor seed.size shriveling
## 562           1             1         1          1    0
## 13            0             0         0          0   13
## 55            0             0         0          0   19
## 15            0             0         0          0   28
## 1             0             0         0          0   30
##              84            84        84         84 1664
```

```
nrow(no_nulls) #646
```

```
## [1] 646
```

```
colSums(is.na(no_nulls)) # Max number of nulls in a column is 84.
```

```
##           Class            date     plant.stand          precip            temp
##               0               1              16              16              16
##            hail       crop.hist        area.dam           sever        seed.tmt
##              84              16               1              84              84
##            germ    plant.growth          leaves       leaf.halo       leaf.marg
##              84              16               0              55              55
##       leaf.size     leaf.shread       leaf.malf       leaf.mild            stem
##              55              71              55              71              16
##         lodging    stem.cankers   canker.lesion fruiting.bodies       ext.decay
##              84              16              16              84              16
##        mycelium    int.discolor       sclerotia      fruit.pods     fruit.spots
##              16              16              16              84              84
##            seed     mold.growth   seed.discolor       seed.size      shriveling
##              84              84              84              84              84
##           roots
##              16
```