DATA624-HW4-Preprocessing

Kuhn & Johnson exercises 3.1, 3.2

Michael Y.

3/1/2020

Contents

```r
library(fpp2)
```

```
## Loading required package: ggplot2

## Loading required package: forecast

## Registered S3 method overwritten by 'xts':
##   method     from
##   as.zoo.xts zoo

## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo

## Registered S3 methods overwritten by 'forecast':
##   method             from
##   fitted.fracdiff    fracdiff
##   residuals.fracdiff fracdiff

## Loading required package: fma

## Loading required package: expsmooth
```

```r
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------------------- tidyverse 1.3.0 --

## <U+2713> tibble  2.1.3     <U+2713> dplyr   0.8.3
## <U+2713> tidyr   1.0.0     <U+2713> stringr 1.4.0
## <U+2713> readr   1.3.1     <U+2713> forcats 0.4.0
## <U+2713> purrr   0.3.3

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

```r
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```r
library(e1071)
library(caret)
```

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

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```r
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```r
library(AppliedPredictiveModeling)
```

Homework 4 - Preprocessing

Do problems 3.1 and 3.2 in the Kuhn and Johnson book Applied Predictive Modeling. Please submit both your Rpubs link as well as attach the .rmd file with your code.

3.1 Glass identification

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 ("RI") and
- percentages of eight elements:
  - Na, Mg, Al, Si, K, Ca, Ba, and Fe.

The data can be accessed via:

```
library(mlbench)
data(Glass)
str(Glass)
```

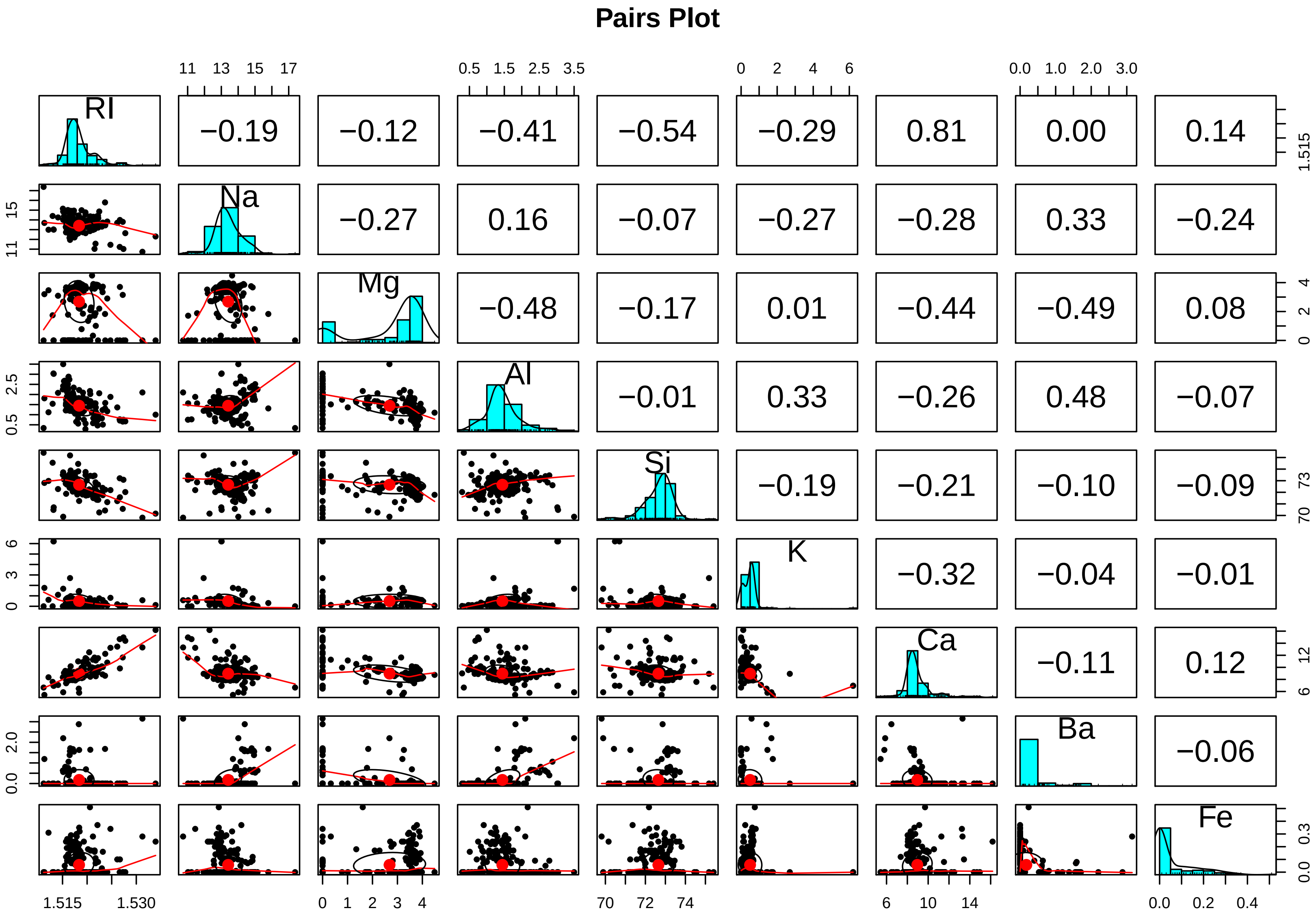
```
## 'data.frame':    214 obs. of  10 variables:
##  $ RI  : num  1.52 1.52 1.52 1.52 1.52 ...
##  $ Na  : num  13.6 13.9 13.5 13.2 13.3 ...
##  $ Mg  : num  4.49 3.6 3.55 3.69 3.62 3.61 3.6 3.61 3.58 3.6 ...
##  $ Al  : num  1.1 1.36 1.54 1.29 1.24 1.62 1.14 1.05 1.37 1.36 ...
##  $ Si  : num  71.8 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 ...
```

```
#'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 ...
```

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

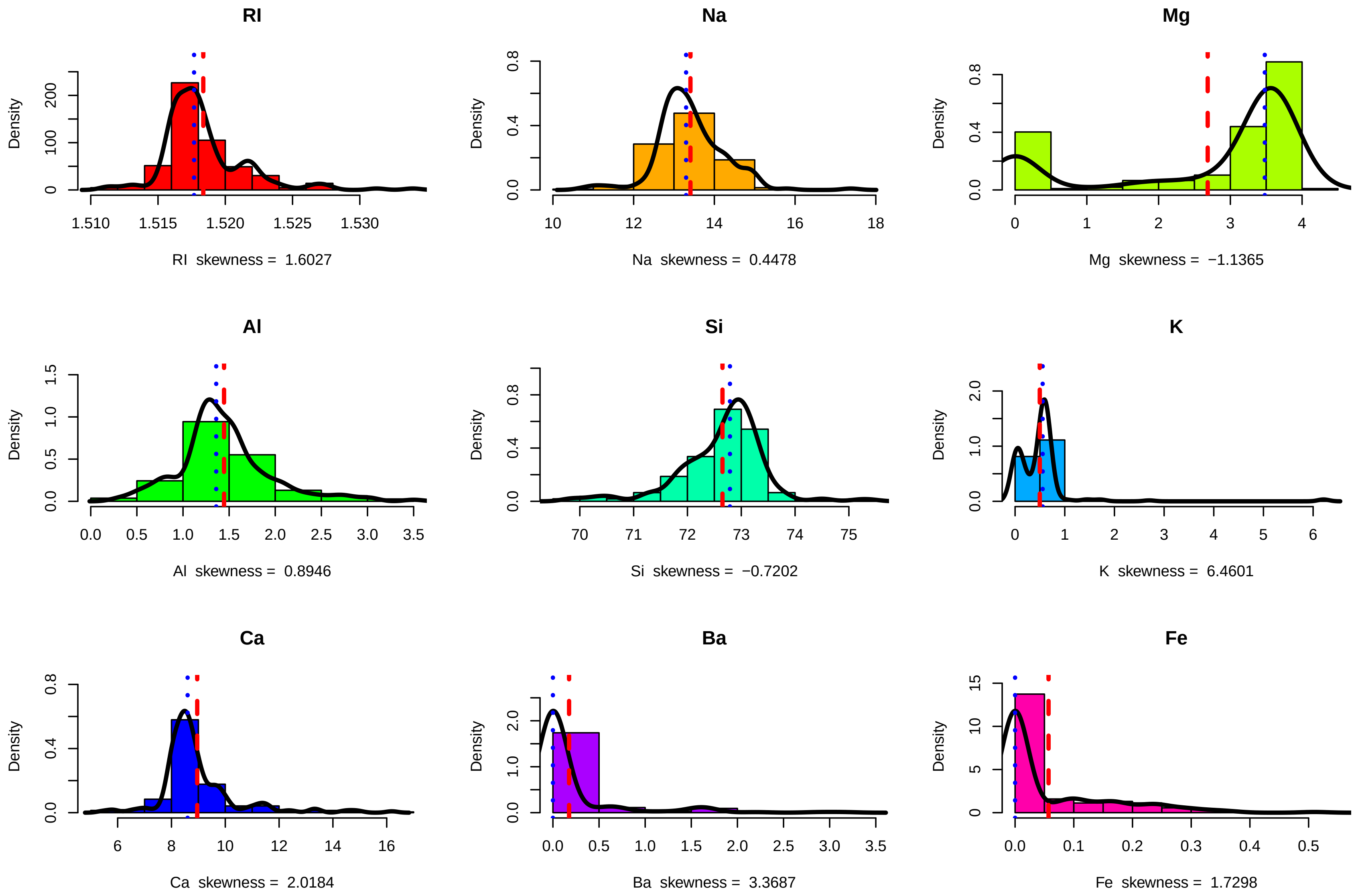
```r
Glass %>%
  select (-Type) -> Glass_Predictors
```

```r
#### Pairs plot
psych::pairs.panels(Glass_Predictors, main="Pairs Plot")
```


Pairs Plot
11 13 15 17
0.5 1.5 2.5 3.5
0 2 4 6
0.0 1.0 2.0 3.0
RI
−0.19
−0.12
−0.41
−0.54
−0.29
0.81
0.00
0.14
1.515
15
11
Na
−0.27
0.16
−0.07
−0.27
−0.28
0.33
−0.24
Mg
−0.48
−0.17
0.01
−0.44
−0.49
0.08
4
2
0
2.5
0.5
Al
−0.01
0.33
−0.26
0.48
−0.07
Si
−0.19
−0.21
−0.10
−0.09
73
70
6
3
0
K
−0.32
−0.04
−0.01
Ca
−0.11
0.12
12
6
2.0
0.0
Ba
−0.06
Fe
0.3
0.0
1.515 1.530
0 1 2 3 4
70 72 74
6 10 14
0.0 0.2 0.4

```
#### Histogram and Density
par(mfrow=c(3,3),oma= c(0, 0, 2, 0))
for (col in 1:ncol(Glass_Predictors)) {
    hist(Glass_Predictors[,col],
         col=rainbow(9)[col],
         prob=TRUE,
         main=names(Glass_Predictors[col]),
         xlab=paste(names(Glass_Predictors[col])," skewness = ",
                    round(skewness(Glass_Predictors[,col]),4)),
         ylim=c(0,1.3*max(density(Glass_Predictors[,col])$y))
         )
    lines(density(Glass_Predictors[,col]),lwd=3)
    abline(v=median(Glass_Predictors[,col]),lwd=3,lty=3, col="blue")
    abline(v=mean(Glass_Predictors[,col]),lwd=3,lty=2, col="red")
mtext("Histogram, Densities, Means, Medians for Glass Components",
      side = 3, line = +0.5, outer = TRUE, cex=1.5)

}
```

Histogram, Densities, Means, Medians for Glass Components


RI

Na

Mg

Al

Si

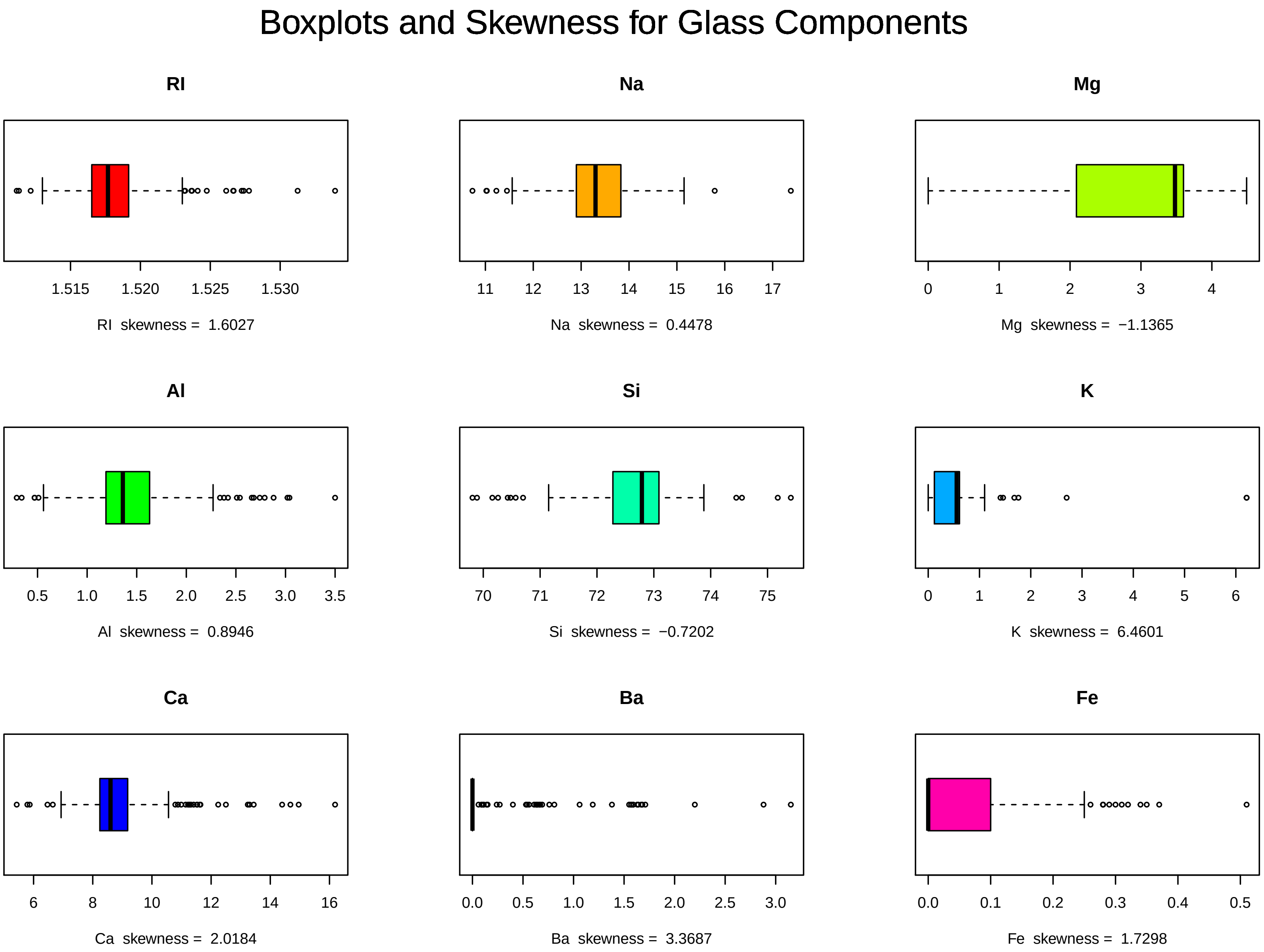
K

Ca

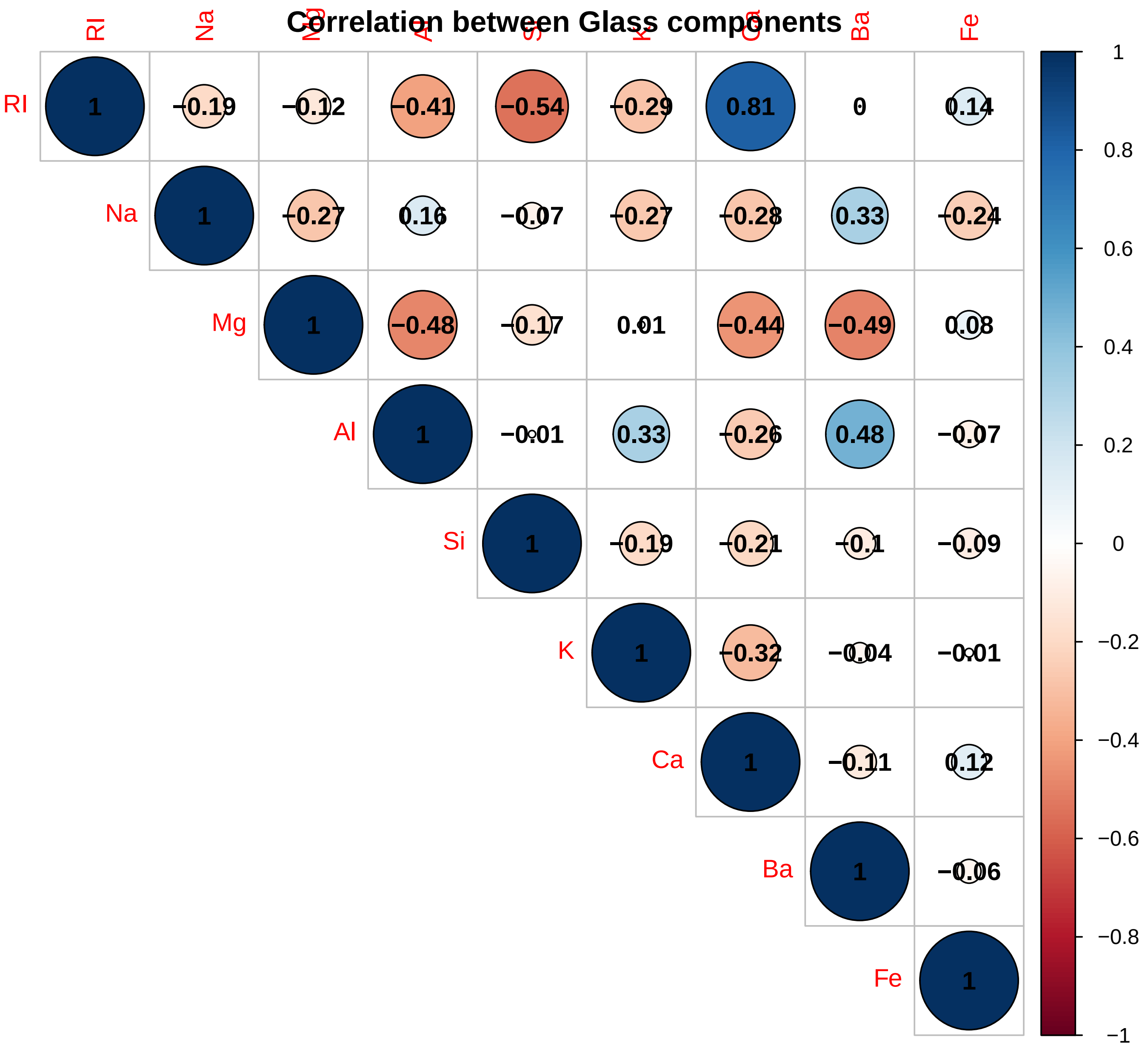
Ba

Fe

```
#### Boxplots
par(mfrow=c(3,3),oma= c(0, 0, 2, 0))
for (col in 1:ncol(Glass_Predictors)) {
    boxplot(Glass_Predictors[,col],
         col=rainbow(9)[col],
         horizontal = TRUE,
         main=names(Glass_Predictors[col]),
         xlab=paste(names(Glass_Predictors[col])," skewness = ",
                    round(skewness(Glass_Predictors[,col]),4))
         )

mtext("Boxplots and Skewness for Glass Components",
      side = 3, line = +0.5, outer = TRUE, cex=1.5)

}
```


Boxplots and Skewness for Glass Components
RI
1.515
1.520
1.525
1.530
RI skewness = 1.6027
Na
11
12
13
14
15
16
17
Na skewness = 0.4478
Mg
0
1
2
3
4
Mg skewness = −1.1365
Al
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Al skewness = 0.8946
Si
70
71
72
73
74
75
Si skewness = −0.7202
K
0
1
2
3
4
5
6
K skewness = 6.4601
Ca
6
8
10
12
14
16
Ca skewness = 2.0184
Ba
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Ba skewness = 3.3687
Fe
0.0
0.1
0.2
0.3
0.4
0.5
Fe skewness = 1.7298

```
#### Correlations
GlassCorr <- cor(Glass_Predictors)

corrplot(corr = GlassCorr, type = "upper", outline = T, order="original",
           sig.level = 0.05, insig = "blank", addCoef.col = "black",
           title = "\nCorrelation between Glass components",
           number.cex = 1.0, number.font = 2, number.digits = 2 )
```


Correlation between Glass components

There is a very high positive correlation, +.81, between `Ca` and the Refractive Index, `RI`.
There are a number of moderately strong correlations, both positive and negative, with values close to ± 0.5.

Table 1: Skewness

	x
Mg	-1.1364523
Si	-0.7202392
Na	0.4478343
Al	0.8946104
RI	1.6027151
Fe	1.7298107
Ca	2.0184463
Ba	3.3686800
K	6.4600889

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

```
# Skewness
apply(X = Glass_Predictors, MARGIN = 2, FUN=skewness) %>% sort -> Glass_Skew
Glass_Skew %>%
  kable(caption = "Skewness") %>%
  kable_styling(c("bordered","striped"),full_width = F)
```

The boxplots reveal numerous outliers, most notably with **K**.
Because most of the values for **Ba** (176) and **Fw** (144) are zeroes, all the other points for these elements appear as "outliers" under the standard boxplot method.

Mg is heavily skewed to the **left**.
K, Ba, Ca, and **Fe** are heavily skewed to the **right**.
RI and **Al** are mildly skewed to the **right**.

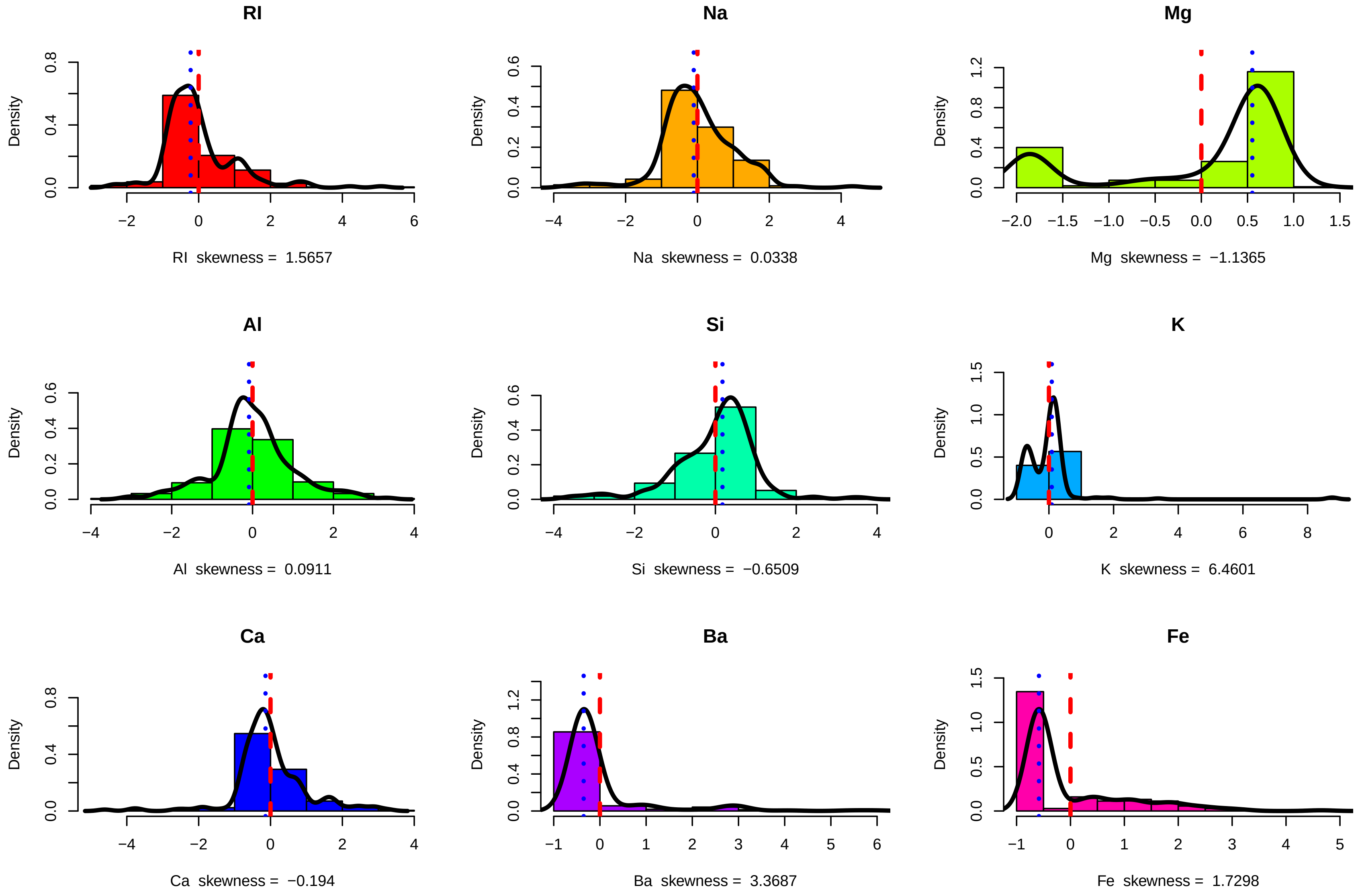
(c) Are there any relevant transformations of one or more predictors that might improve the classification model?

We'll try the following set of transformations:

- Box-Cox transformation
- Center the variables at mean=0
- Scale to stdev=1

```
Glass_BoxCox1 <- predict(preProcess(Glass_Predictors,
                                    method=c('BoxCox','center','scale')),
                         Glass_Predictors)

par(mfrow=c(3,3),oma= c(0, 0, 2, 0))
for (col in 1:ncol(Glass_BoxCox1)) {
    hist(Glass_BoxCox1[,col],
         col=rainbow(9)[col],
         prob=TRUE,
         main=names(Glass_BoxCox1[col]),
         xlab=paste(names(Glass_BoxCox1[col])," skewness = ",
                    round(skewness(Glass_BoxCox1[,col]),4)),
         ylim=c(0,1.3*max(density(Glass_BoxCox1[,col])$y))
         )
    lines(density(Glass_BoxCox1[,col]),lwd=3)
    abline(v=median(Glass_BoxCox1[,col]),lwd=3,lty=3, col="blue")
    abline(v=mean(Glass_BoxCox1[,col]),lwd=3,lty=2, col="red")
mtext("Histogram, Densities, for BoxCox, center, and scale transforms",
      side = 3, line = +0.5, outer = TRUE, cex=1.5)
}
```

Histogram, Densities, for BoxCox, center, and scale transforms



```r
#### Boxplots following transformations
par(mfrow=c(3,3),oma= c(0, 0, 2, 0))
for (col in 1:ncol(Glass_BoxCox1)) {
    boxplot(Glass_BoxCox1[,col],
         col=rainbow(9)[col],
         horizontal = TRUE,
         main=names(Glass_BoxCox1[col]),
         xlab=paste(names(Glass_BoxCox1[col])," skewness = ",
                    round(skewness(Glass_BoxCox1[,col]),4))
         )

mtext("Boxplots and Skewness for transformed Glass Components",
      side = 3, line = +0.5, outer = TRUE, cex=1.5)

}
```

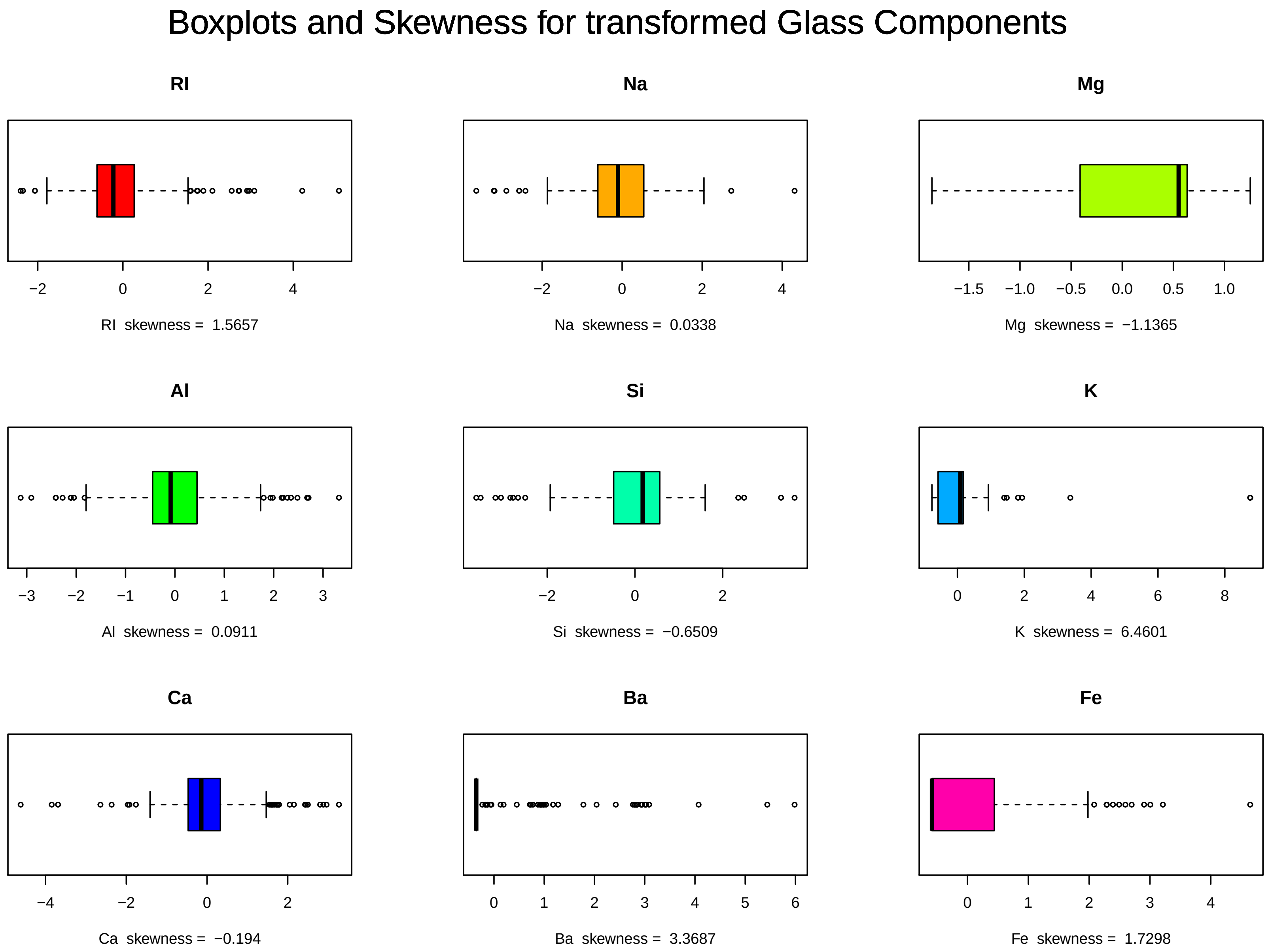

Boxplots and Skewness for transformed Glass Components
RI
−2
0
2
4
RI skewness = 1.5657
Na
−2
0
2
4
Na skewness = 0.0338
Mg
−1.5
−1.0
−0.5
0.0
0.5
1.0
Mg skewness = −1.1365
Al
−3
−2
−1
0
1
2
3
Al skewness = 0.0911
Si
−2
0
2
Si skewness = −0.6509
K
0
2
4
6
8
K skewness = 6.4601
Ca
−4
−2
0
2
Ca skewness = −0.194
Ba
0
1
2
3
4
5
6
Ba skewness = 3.3687
Fe
0
1
2
3
4
Fe skewness = 1.7298

Table 2: Skewness, before and after transformation

	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
orig	1.602715	0.4478343	-1.136452	0.8946104	-0.7202392	6.460089	2.0184463	3.36868	1.729811
xformed	1.565660	0.0338464	-1.136452	0.0910590	-0.6509057	6.460089	-0.1939557	3.36868	1.729811

The variables are now all standardized with a centered mean=0 and stdev=1. The Box-Cox transformation does not impact most of these variables. It does improve the following items:

- **Na**, for which the skewness has reduced from 0.4478 to 0.0338
- **Al**, for which the skewness has reduced from 0.8946 to 0.0911
- **Ca**, for which the skewness has reduced from 2.0184 to -0.194

The Box-Cox transformation doesn't have an appreciable improvement on the skewness of the other variables.

```r
rbind(orig=apply(X = Glass_Predictors, MARGIN = 2, FUN=skewness),
      xformed=apply(X = Glass_BoxCox1, MARGIN = 2, FUN=skewness)) %>%
  kable(caption = "Skewness, before and after transformation")%>%
    kable_styling(c("bordered","striped"))
```

Skewness before and after Box-Cox transformation

3.2 Diseased Soybeans

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:

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

Description

There are 19 classes, only the first 15 of which have been used in prior work.
The folklore seems to be that the last four classes are unjustified by the data since they have so few examples.
There are 35 categorical attributes, some nominal and some ordered.
The value "dna" means does not apply.
The values for attributes are encoded numerically, with the first value encoded as "0," the second as "1," and so forth.

Format

A data frame with 683 observations on 36 variables.
There are 35 categorical attributes, all numerical and a nominal denoting the class.

[,1] `Class` the 19 classes
[,2] `date` apr(0),may(1),june(2),july(3),aug(4),sept(5),oct(6).
[,3] `plant.stand` normal(0),lt-normal(1).
[,4] `precip` lt-norm(0),norm(1),gt-norm(2).
[,5] `temp` lt-norm(0),norm(1),gt-norm(2).
[,6] `hail` yes(0),no(1).
[,7] `crop.hist` dif-lst-yr(0),s-l-y(1),s-l-2-y(2), s-l-7-y(3).
[,8] `area.dam` scatter(0),low-area(1),upper-ar(2),whole-field(3).
[,9] `sever` minor(0),pot-severe(1),severe(2).
[,10] `seed.tmt` none(0),fungicide(1),other(2).
[,11] `germ` 90-100%(0),80-89%(1),lt-80%(2).
[,12] `plant.growth` norm(0),abnorm(1).
[,13] `leaves` norm(0),abnorm(1).
[,14] `leaf.halo` absent(0),yellow-halos(1),no-yellow-halos(2).

[,15] `leaf.marg` w-s-marg(0),no-w-s-marg(1),dna(2).

[,16] `leaf.size` lt-1/8(0),gt-1/8(1),dna(2).

[,17] `leaf.shread` absent(0),present(1).

[,18] `leaf.malf` absent(0),present(1).

[,19] `leaf.mild` absent(0),upper-surf(1),lower-surf(2).

[,20] `stem` norm(0),abnorm(1).

[,21] ‘`lodging` yes(0),no(1).

[,22] `stem.cankers` absent(0),below-soil(1),above-s(2),ab-sec-nde(3).

[,23] `canker.lesion` dna(0),brown(1),dk-brown-blk(2),tan(3).

[,24] `fruiting.bodies` absent(0),present(1).

[,25] `ext.decay` absent(0),firm-and-dry(1),watery(2).

[,26] `mycelium` absent(0),present(1).

[,27] `int.discolor` none(0),brown(1),black(2).

[,28] `sclerotia` absent(0),present(1).

[,29] `fruit.pods` norm(0),diseased(1),few-present(2),dna(3).

[,30] `fruit.spots` absent(0),col(1),br-w/blk-speck(2),distort(3),dna(4).

[,31] `seed` norm(0),abnorm(1).

[,32] `mold.growth` absent(0),present(1).

[,33] `seed.discolor` absent(0),present(1).

[,34] `seed.size` norm(0),lt-norm(1).

[,35] `shriveling` absent(0),present(1).

[,36] `roots` norm(0),rotted(1),galls-cysts(2).

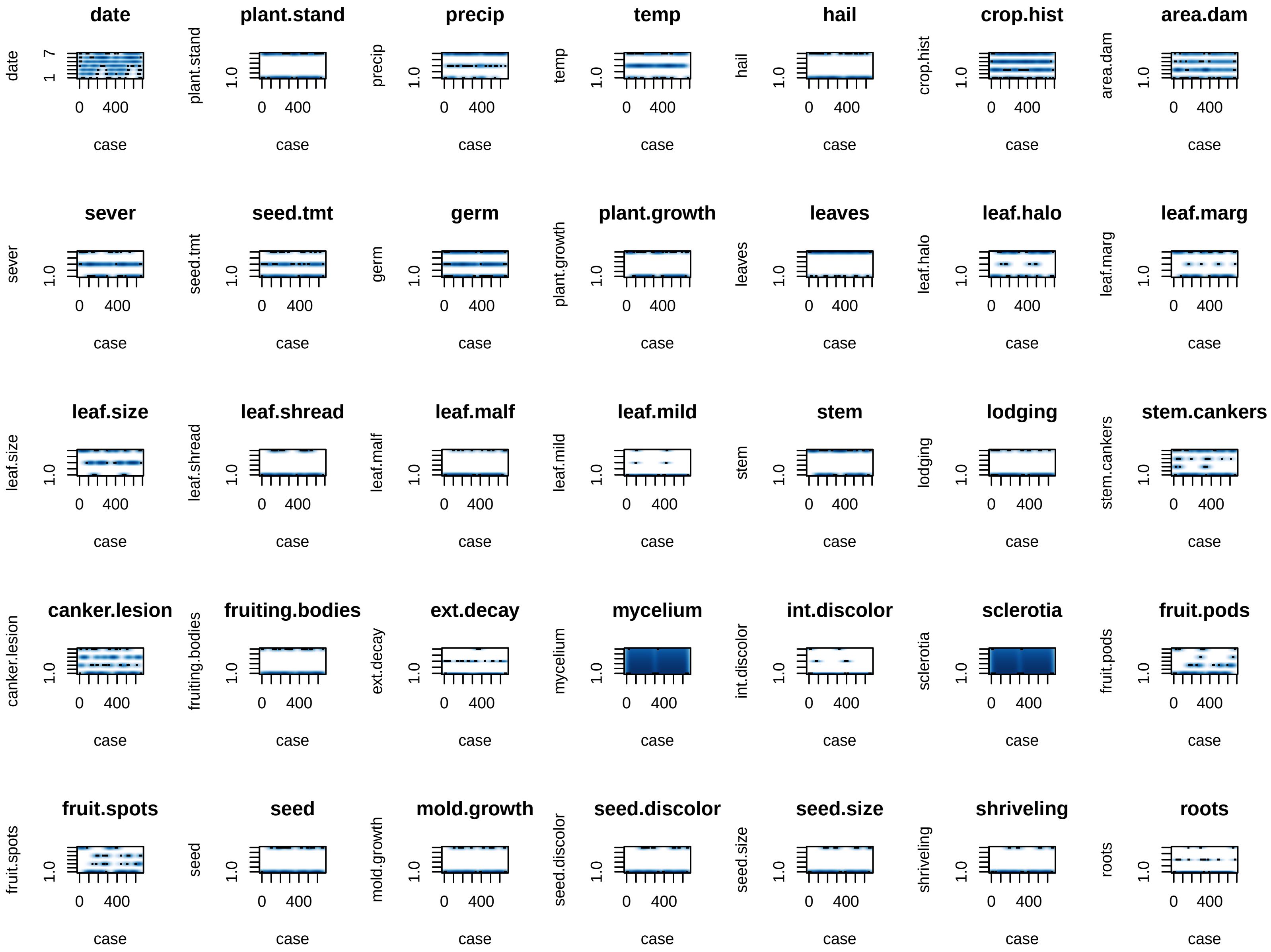
(a) Investigate the frequency distributions for the categorical predictors.

```
### Summary
summary(Soybean)
```

```
##                   Class          date     plant.stand  precip      temp    
##  brown-spot         : 92   5      :149   0   :354   0   : 74   0   : 80  
##  alternarialeaf-spot: 91   4      :131   1   :293   1   :112   1   :374  
##  frog-eye-leaf-spot : 91   3      :118   NA's: 36   2   :459   2   :199  
##  phytophthora-rot   : 88   2      : 93              NA's: 38   NA's: 30  
##  anthracnose        : 44   6      : 90                                   
##  brown-stem-rot     : 44   (Other):101                                   
##  (Other)            :233   NA's   :  1                                   
##    hail     crop.hist  area.dam    sever     seed.tmt     germ     plant.growth
##  0   :435   0   : 65   0   :123   0   :195   0   :305   0   :165   0   :441   
##  1   :127   1   :165   1   :227   1   :322   1   :222   1   :213   1   :226   
##  NA's:121   2   :219   2   :145   2   : 45   2   : 35   2   :193   NA's: 16   
##             3   :218   3   :187   NA's:121   NA's:121   NA's:112              
##             NA's: 16   NA's:  1                                               
##                                                                               
##                                                                               
##  leaves  leaf.halo  leaf.marg  leaf.size  leaf.shread leaf.malf  leaf.mild 
##  0: 77   0   :221   0   :357   0   : 51   0   :487    0   :554   0   :535  
##  1:606   1   : 36   1   : 21   1   :327   1   : 96    1   : 45   1   : 20  
##          2   :342   2   :221   2   :221   NA's:100    NA's: 84   2   : 20  
##          NA's: 84   NA's: 84   NA's: 84                          NA's:108  
##                                                                            
##                                                                            
##                                                                            
##    stem     lodging    stem.cankers canker.lesion fruiting.bodies ext.decay 
##  0   :296   0   :520   0   :379     0   :320      0   :473        0   :497  
##  1   :371   1   : 42   1   : 39     1   : 83      1   :104        1   :135  
##  NA's: 16   NA's:121   2   : 36     2   :177      NA's:106        2   : 13  
##                        3   :191     3   : 65                      NA's: 38  
##                        NA's: 38     NA's: 38                                
##                                                                             
##                                                                             
##  mycelium   int.discolor sclerotia  fruit.pods fruit.spots   seed    
##  0   :639   0   :581     0   :625   0   :407   0   :345    0   :476
##  1   :  6   1   : 44     1   : 20   1   :130   1   : 75    1   :115
##  NA's: 38   2   : 20     NA's: 38   2   : 14   2   : 57    NA's: 92
##             NA's: 38                3   : 48   4   :100
##                                     NA's: 84   NA's:106
##
##
##  mold.growth seed.discolor seed.size  shriveling  roots
##  0   :524    0   :513      0   :532   0   :539   0   :551
##  1   : 67    1   : 64      1   : 59   1   : 38   1   : 86
##  NA's: 92    NA's:106      NA's: 92   NA's:106   2   : 15
##                                                  NA's: 31
##
##
##
```

```
n = ncol(Soybean)
```

```
SoybeanPred <- Soybean[,-1]
### Scatterplots
par(mfrow = c(5,7))
for (i in 1:ncol(SoybeanPred)) {
  smoothScatter(SoybeanPred[ ,i],
                xlab="case",
                ylab =names(SoybeanPred[i]),
                main = names(SoybeanPred[i]))
}
```


date
date
1
7
0
400
case
plant.stand
plant.stand
1.0
0
400
case
precip
precip
1.0
0
400
case
temp
temp
1.0
0
400
case
hail
hail
1.0
0
400
case
crop.hist
crop.hist
1.0
0
400
case
area.dam
area.dam
1.0
0
400
case
sever
sever
1.0
0
400
case
seed.tmt
seed.tmt
1.0
0
400
case
germ
germ
1.0
0
400
case
plant.growth
plant.growth
1.0
0
400
case
leaves
leaves
1.0
0
400
case
leaf.halo
leaf.halo
1.0
0
400
case
leaf.marg
leaf.marg
1.0
0
400
case
leaf.size
leaf.size
1.0
0
400
case
leaf.shread
leaf.shread
1.0
0
400
case
leaf.malf
leaf.malf
1.0
0
400
case
leaf.mild
leaf.mild
1.0
0
400
case
stem
stem
1.0
0
400
case
lodging
lodging
1.0
0
400
case
stem.cankers
stem.cankers
1.0
0
400
case
canker.lesion
canker.lesion
1.0
0
400
case
fruiting.bodies
fruiting.bodies
1.0
0
400
case
ext.decay
ext.decay
1.0
0
400
case
mycelium
mycelium
1.0
0
400
case
int.discolor
int.discolor
1.0
0
400
case
sclerotia
sclerotia
1.0
0
400
case
fruit.pods
fruit.pods
1.0
0
400
case
fruit.spots
fruit.spots
1.0
0
400
case
seed
seed
1.0
0
400
case
mold.growth
mold.growth
1.0
0
400
case
seed.discolor
seed.discolor
1.0
0
400
case
seed.size
seed.size
1.0
0
400
case
shriveling
shriveling
1.0
0
400
case
roots
roots
1.0
0
400
case

```r
ratio = NULL
tabl = list()
sumtabl = list()
for (i in colnames(Soybean)) {
  print("______________________________________________________________________________")
  print(i);
  tabl[[i]] = sort(table(Soybean[,i]),decreasing = T)
  sumtabl[[i]] = sum(tabl[[i]])
  result=tabl[[i]] / sumtabl[[i]]
  print(result)
    ratio[i] = result[1]/result[2]
  print(paste("ratio of first 2 values in ",i,": ", round(ratio[i],4)))
}
```

```
## [1] "______________________________________________________________________________"
## [1] "Class"
## 
##                  brown-spot         alternarialeaf-spot 
##                  0.13469985                  0.13323572 
##          frog-eye-leaf-spot            phytophthora-rot 
##                  0.13323572                  0.12884334 
##                 anthracnose              brown-stem-rot 
##                  0.06442167                  0.06442167 
##            bacterial-blight           bacterial-pustule 
##                  0.02928258                  0.02928258 
##                charcoal-rot       diaporthe-stem-canker 
##                  0.02928258                  0.02928258 
##                downy-mildew      phyllosticta-leaf-spot 
##                  0.02928258                  0.02928258 
##              powdery-mildew           purple-seed-stain 
##                  0.02928258                  0.02928258 
##        rhizoctonia-root-rot                2-4-d-injury 
##                  0.02928258                  0.02342606 
## diaporthe-pod-&-stem-blight               cyst-nematode 
##                  0.02196193                  0.02049780 
##            herbicide-injury 
##                  0.01171303 
## [1] "ratio of first 2 values in  Class :  1.011"
## [1] "_______________________________________________________________________________"
## [1] "date"
##
##          5          4          3          2          6          1          0
## 0.21847507 0.19208211 0.17302053 0.13636364 0.13196481 0.10997067 0.03812317
## [1] "ratio of first 2 values in  date :  1.1374"
## [1] "_______________________________________________________________________________"
## [1] "plant.stand"
##
##         0         1
## 0.5471406 0.4528594
## [1] "ratio of first 2 values in  plant.stand :  1.2082"
## [1] "_______________________________________________________________________________"
## [1] "precip"
##
##         2         1         0
## 0.7116279 0.1736434 0.1147287
## [1] "ratio of first 2 values in  precip :  4.0982"
## [1] "_______________________________________________________________________________"
## [1] "temp"
##
##         1         2         0
## 0.5727412 0.3047473 0.1225115
## [1] "ratio of first 2 values in  temp :  1.8794"
## [1] "_______________________________________________________________________________"
## [1] "hail"
##
##         0         1
## 0.7740214 0.2259786
## [1] "ratio of first 2 values in  hail :  3.4252"
## [1] "_______________________________________________________________________________"
## [1] "crop.hist"
##
##          2          3          1          0
## 0.32833583 0.32683658 0.24737631 0.09745127
## [1] "ratio of first 2 values in  crop.hist :  1.0046"
## [1] "_______________________________________________________________________________"
## [1] "area.dam"
##
```

```
##         1         3         2         0
## 0.3328446 0.2741935 0.2126100 0.1803519
## [1] "ratio of first 2 values in  area.dam :  1.2139"
## [1] "_______________________________________________________________________________"
## [1] "sever"
##
##          1          0          2
## 0.57295374 0.34697509 0.08007117
## [1] "ratio of first 2 values in  sever :  1.6513"
## [1] "_______________________________________________________________________________"
## [1] "seed.tmt"
##
##          0          1          2
## 0.54270463 0.39501779 0.06227758
## [1] "ratio of first 2 values in  seed.tmt :  1.3739"
## [1] "_______________________________________________________________________________"
## [1] "germ"
##
##         1         2         0
## 0.3730298 0.3380035 0.2889667
## [1] "ratio of first 2 values in  germ :  1.1036"
## [1] "_______________________________________________________________________________"
## [1] "plant.growth"
##
##         0         1
## 0.6611694 0.3388306
## [1] "ratio of first 2 values in  plant.growth :  1.9513"
## [1] "_______________________________________________________________________________"
## [1] "leaves"
##
##         1         0
## 0.8872621 0.1127379
## [1] "ratio of first 2 values in  leaves :  7.8701"
## [1] "_______________________________________________________________________________"
## [1] "leaf.halo"
##
##          2          0          1
## 0.57095159 0.36894825 0.06010017
## [1] "ratio of first 2 values in  leaf.halo :  1.5475"
## [1] "_______________________________________________________________________________"
## [1] "leaf.marg"
##
##          0          2          1
## 0.59599332 0.36894825 0.03505843
## [1] "ratio of first 2 values in  leaf.marg :  1.6154"
## [1] "_______________________________________________________________________________"
## [1] "leaf.size"
##
##         1         2         0
## 0.5459098 0.3689482 0.0851419
## [1] "ratio of first 2 values in  leaf.size :  1.4796"
## [1] "_______________________________________________________________________________"
## [1] "leaf.shread"
##
##         0         1
## 0.8353345 0.1646655
## [1] "ratio of first 2 values in  leaf.shread :  5.0729"
## [1] "_______________________________________________________________________________"
## [1] "leaf.malf"
##
##          0          1
## 0.92487479 0.07512521
## [1] "ratio of first 2 values in  leaf.malf :  12.3111"
## [1] "_______________________________________________________________________________"
## [1] "leaf.mild"
##
##          0          1          2
## 0.93043478 0.03478261 0.03478261
## [1] "ratio of first 2 values in  leaf.mild :  26.75"
## [1] "_______________________________________________________________________________"
## [1] "stem"
##
##         1         0
## 0.5562219 0.4437781
## [1] "ratio of first 2 values in  stem :  1.2534"
## [1] "_______________________________________________________________________________"
## [1] "lodging"
##
```

```
##         0         1
## 0.9252669 0.0747331
## [1] "ratio of first 2 values in  lodging :  12.381"
## [1] "_______________________________________________________________________________"
## [1] "stem.cankers"
##
##          0          3          1          2
## 0.58759690 0.29612403 0.06046512 0.05581395
## [1] "ratio of first 2 values in  stem.cankers :  1.9843"
## [1] "_______________________________________________________________________________"
## [1] "canker.lesion"
##
##         0         2         1         3
## 0.4961240 0.2744186 0.1286822 0.1007752
## [1] "ratio of first 2 values in  canker.lesion :  1.8079"
## [1] "_______________________________________________________________________________"
## [1] "fruiting.bodies"
##
##         0         1
## 0.8197574 0.1802426
## [1] "ratio of first 2 values in  fruiting.bodies :  4.5481"
## [1] "_______________________________________________________________________________"
## [1] "ext.decay"
##
##          0          1          2
## 0.77054264 0.20930233 0.02015504
## [1] "ratio of first 2 values in  ext.decay :  3.6815"
## [1] "_______________________________________________________________________________"
## [1] "mycelium"
##
##           0           1
## 0.990697674 0.009302326
## [1] "ratio of first 2 values in  mycelium :  106.5"
## [1] "_______________________________________________________________________________"
## [1] "int.discolor"
##
##          0          1          2
## 0.90077519 0.06821705 0.03100775
## [1] "ratio of first 2 values in  int.discolor :  13.2045"
## [1] "_______________________________________________________________________________"
## [1] "sclerotia"
##
##          0          1
## 0.96899225 0.03100775
## [1] "ratio of first 2 values in  sclerotia :  31.25"
## [1] "_______________________________________________________________________________"
## [1] "fruit.pods"
##
##          0          1          3          2
## 0.67946578 0.21702838 0.08013356 0.02337229
## [1] "ratio of first 2 values in  fruit.pods :  3.1308"
## [1] "_______________________________________________________________________________"
## [1] "fruit.spots"
##
##          0          4          1          2
## 0.59792028 0.17331023 0.12998267 0.09878683
## [1] "ratio of first 2 values in  fruit.spots :  3.45"
## [1] "_______________________________________________________________________________"
## [1] "seed"
##
##         0         1
## 0.8054146 0.1945854
## [1] "ratio of first 2 values in  seed :  4.1391"
## [1] "_______________________________________________________________________________"
## [1] "mold.growth"
##
##         0         1
## 0.8866328 0.1133672
## [1] "ratio of first 2 values in  mold.growth :  7.8209"
## [1] "_______________________________________________________________________________"
## [1] "seed.discolor"
##
##         0         1
## 0.8890815 0.1109185
## [1] "ratio of first 2 values in  seed.discolor :  8.0156"
## [1] "_______________________________________________________________________________"
## [1] "seed.size"
##
```

```
##         0         1
## 0.9001692 0.0998308
## [1] "ratio of first 2 values in  seed.size :  9.0169"
## [1] "_______________________________________________________________________________"
## [1] "shriveling"
##
##          0          1
## 0.93414211 0.06585789
## [1] "ratio of first 2 values in  shriveling :  14.1842"
## [1] "_______________________________________________________________________________"
## [1] "roots"
##
##          0          1          2
## 0.84509202 0.13190184 0.02300613
## [1] "ratio of first 2 values in  roots :  6.407"
```

```r
# compute the variance on each column (excluding the first column, Class), ignoring NAs
apply(X=Soybean[,2:36],MARGIN = 2, FUN=var, na.rm=T) %>%
  sort() %>%
  kable(caption="Variance of each predictor") %>%
  kable_styling(c("bordered","striped"),full_width = F)
```

Variance of each predictor

Are any of the distributions degenerate in the ways discussed earlier in this chapter? Here are the definitions from the book:

zero variance predictor: a predictor variable that has a single unique value There are no such variables with zero variance.

near-zero variance predictors:

- may have a single value for the vast majority of the samples;
- some predictors might have only a handful of unique values that occur with very low frequencies

Rule-of-thumb: The fraction of unique values over the sample size is low (say 10%).

All variables meet this criterion.

The ratio of the frequency of the most prevalent value to the frequency of the second most prevalent value is large (say around 20).

The following variables meet this criterion:

- leaf.mild
- mycelium
- sclerotia

Table 3: Variance of each predictor

	x
mycelium	0.0092301
sclerotia	0.0300929
shriveling	0.0616274
lodging	0.0692713
leaf.malf	0.0695976
seed.size	0.0900169
seed.discolor	0.0987868
leaves	0.1001748
mold.growth	0.1006854
leaf.shread	0.1377871
fruiting.bodies	0.1480117
seed	0.1569876
leaf.mild	0.1633086
hail	0.1752241
int.discolor	0.1755597
roots	0.1925683
plant.growth	0.2243608
ext.decay	0.2279696
stem	0.2472097
plant.stand	0.2481613
sever	0.3564428
leaf.size	0.3741688
seed.tmt	0.3748390
temp	0.3946533
precip	0.4707978
germ	0.6256614
fruit.pods	0.7788287
leaf.halo	0.9005980
leaf.marg	0.9149195
crop.hist	0.9521185
area.dam	1.1542798
canker.lesion	1.1750590
stem.cankers	1.8270836
fruit.spots	2.2599834
date	2.8700333

There is a function, **caret::nearZeroVar** , which can compute this directly:

```
caret::nearZeroVar(Soybean[,2:36],names=T,saveMetrics = T)
```

```
##                  freqRatio percentUnique zeroVar   nzv
## date              1.137405     1.0248902   FALSE FALSE
## plant.stand       1.208191     0.2928258   FALSE FALSE
## precip            4.098214     0.4392387   FALSE FALSE
## temp              1.879397     0.4392387   FALSE FALSE
## hail              3.425197     0.2928258   FALSE FALSE
## crop.hist         1.004587     0.5856515   FALSE FALSE
## area.dam          1.213904     0.5856515   FALSE FALSE
## sever             1.651282     0.4392387   FALSE FALSE
## seed.tmt          1.373874     0.4392387   FALSE FALSE
## germ              1.103627     0.4392387   FALSE FALSE
## plant.growth      1.951327     0.2928258   FALSE FALSE
## leaves            7.870130     0.2928258   FALSE FALSE
## leaf.halo         1.547511     0.4392387   FALSE FALSE
## leaf.marg         1.615385     0.4392387   FALSE FALSE
## leaf.size         1.479638     0.4392387   FALSE FALSE
## leaf.shread       5.072917     0.2928258   FALSE FALSE
## leaf.malf        12.311111     0.2928258   FALSE FALSE
## leaf.mild        26.750000     0.4392387   FALSE  TRUE
## stem              1.253378     0.2928258   FALSE FALSE
## lodging          12.380952     0.2928258   FALSE FALSE
## stem.cankers      1.984293     0.5856515   FALSE FALSE
## canker.lesion     1.807910     0.5856515   FALSE FALSE
## fruiting.bodies   4.548077     0.2928258   FALSE FALSE
## ext.decay         3.681481     0.4392387   FALSE FALSE
## mycelium        106.500000     0.2928258   FALSE  TRUE
## int.discolor     13.204545     0.4392387   FALSE FALSE
## sclerotia        31.250000     0.2928258   FALSE  TRUE
## fruit.pods        3.130769     0.5856515   FALSE FALSE
## fruit.spots       3.450000     0.5856515   FALSE FALSE
## seed              4.139130     0.2928258   FALSE FALSE
## mold.growth       7.820896     0.2928258   FALSE FALSE
## seed.discolor     8.015625     0.2928258   FALSE FALSE
## seed.size         9.016949     0.2928258   FALSE FALSE
## shriveling       14.184211     0.2928258   FALSE FALSE
## roots              6.406977     0.4392387   FALSE FALSE
```

(b) Roughly 18% of the data are missing.

```
Soybean.incomplete = Soybean[!complete.cases(Soybean),]

# Dimension of Soybean.incomplete
Soybean_incomplete_rows <- nrow(Soybean.incomplete)

Soybean.complete = Soybean[complete.cases(Soybean),]

# Dimension of Soybean.complete
Soybean_complete_rows <- nrow(Soybean.complete)
```

The number of cases which are missing some data is 121 out of 683 total cases.

Table 4: Classes with missing data elements

x
2-4-d-injury
cyst-nematode
diaporthe-pod-&-stem-blight
herbicide-injury
phytophthora-rot

Is the pattern of missing data related to the classes? The missing data occurs in the following classes:

```r
# List of classes with missing elements
Soybean.incomplete$Class%>%
  unique() %>%
  sort() %>%
  kable(caption = "Classes with missing data elements") %>%
  kable_styling(c("bordered","striped"),full_width = F)
```

The missing data is all in 5 cases.

Are there particular predictors that are more likely to be missing?

Which columns have missing data, and what is the pattern for the missing data?

```r
library(VIM)
```

Let’s leverage the VIM package to get this information.

```
## Loading required package: colorspace
```

```
## Loading required package: grid
```

```
## Loading required package: data.table
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

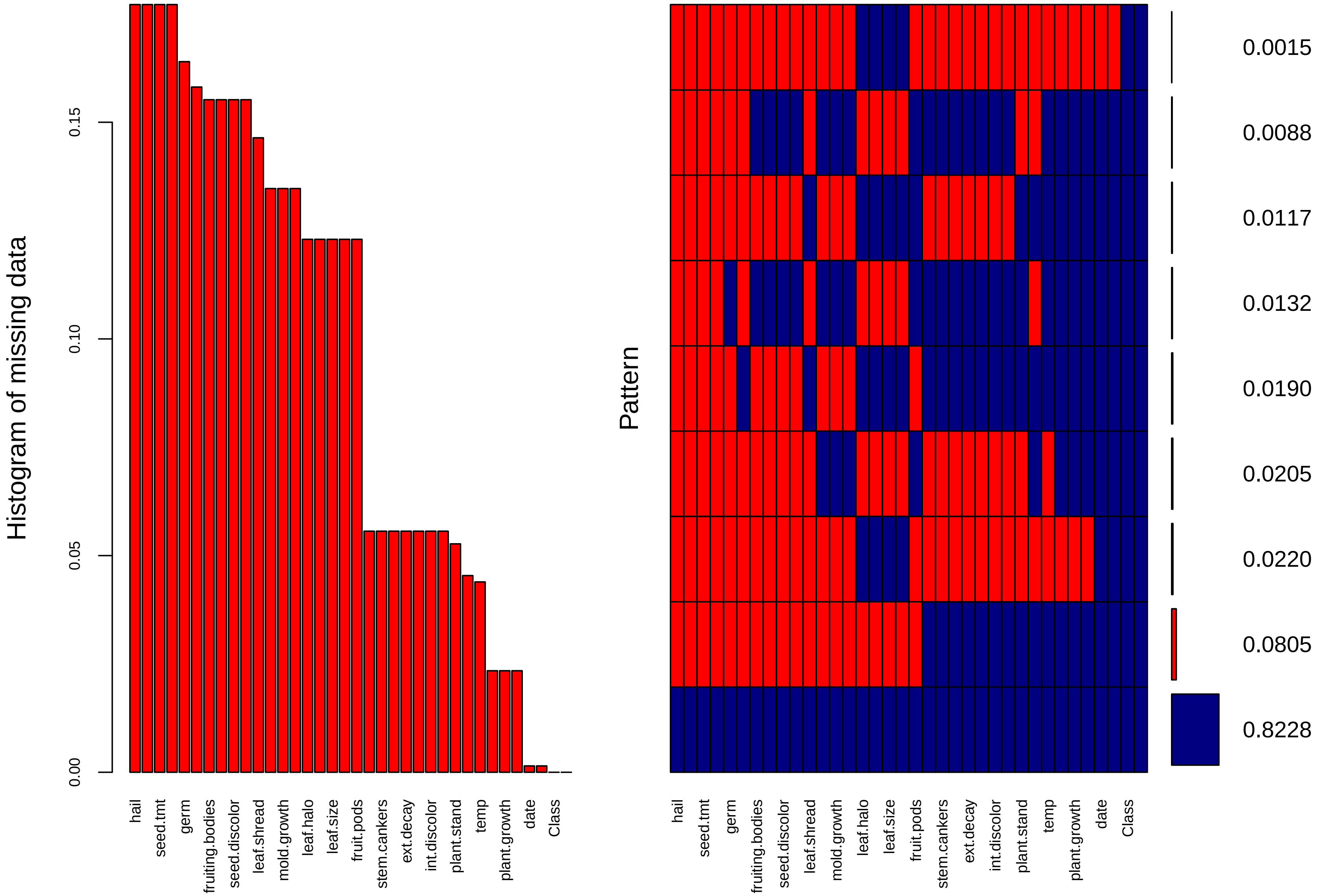
```
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
## VIM is ready to use.
##  Since version 4.0.0 the GUI is in its own package VIMGUI.
##
##           Please use the package to use the new (and old) GUI.
```

```
## Suggestions and bug-reports can be submitted at: https://github.com/alexkowa/VIM/issues
```

```
##
## Attaching package: 'VIM'
## The following object is masked from 'package:datasets':
##
##     sleep
```

```
ggr_plot <- aggr(Soybean, col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE,
                 labels=names(Soybean), cex.axis=.7, gap=3,
                 ylab=c("Histogram of missing data","Pattern"))
```


Histogram of missing data
0.00
0.05
0.10
0.15
hail
seed.tmt
germ
fruiting.bodies
seed.discolor
leaf.shread
mold.growth
leaf.halo
leaf.size
fruit.pods
stem.cankers
ext.decay
int.discolor
plant.stand
temp
plant.growth
date
Class
Pattern
hail
seed.tmt
germ
fruiting.bodies
seed.discolor
leaf.shread
mold.growth
leaf.halo
leaf.size
fruit.pods
stem.cankers
ext.decay
int.discolor
plant.stand
temp
plant.growth
date
Class
0.0015
0.0088
0.0117
0.0132
0.0190
0.0205
0.0220
0.0805
0.8228

```
##
##  Variables sorted by number of missings:
##         Variable       Count
##             hail 0.177159590
##            sever 0.177159590
##         seed.tmt 0.177159590
##          lodging 0.177159590
##             germ 0.163982430
##        leaf.mild 0.158125915
##  fruiting.bodies 0.155197657
##      fruit.spots 0.155197657
##    seed.discolor 0.155197657
##       shriveling 0.155197657
##      leaf.shread 0.146412884
##             seed 0.134699854
##      mold.growth 0.134699854
##        seed.size 0.134699854
##        leaf.halo 0.122986823
##        leaf.marg 0.122986823
##        leaf.size 0.122986823
##        leaf.malf 0.122986823
##       fruit.pods 0.122986823
##           precip 0.055636896
##     stem.cankers 0.055636896
##    canker.lesion 0.055636896
##        ext.decay 0.055636896
##         mycelium 0.055636896
##     int.discolor 0.055636896
##        sclerotia 0.055636896
##      plant.stand 0.052708638
##            roots 0.045387994
##             temp 0.043923865
##        crop.hist 0.023426061
##     plant.growth 0.023426061
##             stem 0.023426061
##             date 0.001464129
##         area.dam 0.001464129
##            Class 0.000000000
##           leaves 0.000000000
```

For the classes which have missing elements, the count of such missing elements is as follows:

```
#Number of missing elements
Soybean.incomplete %>%
  summarize_all(list(
    ~ sum(is.na(.)))
    ) %>%
  sort(decreasing = T) %>%
  t() %>%
  kable(caption="Count of missing elements") %>%
  kable_styling(c("bordered","striped"),full_width = F)
```

For the 5 Classes with missing elements, “hail”, “sever”, “seed.tmt” and “lodging” are entirely absent for all cases. Others elements are missing in accordance with the above table.

Table 5: Count of missing elements

hail	121
sever	121
seed.tmt	121
lodging	121
germ	112
leaf.mild	108
fruiting.bodies	106
fruit.spots	106
seed.discolor	106
shriveling	106
leaf.shread	100
seed	92
mold.growth	92
seed.size	92
leaf.halo	84
leaf.marg	84
leaf.size	84
leaf.malf	84
fruit.pods	84
precip	38
stem.cankers	38
canker.lesion	38
ext.decay	38
mycelium	38
int.discolor	38
sclerotia	38
plant.stand	36
roots	31
temp	30
crop.hist	16
plant.growth	16
stem	16
date	1
area.dam	1
Class	0
leaves	0

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

Let’s use the MICE package to impute missing values MICE: Multivariate Imputation by Chained Equations

```
library(mice)
```

```
## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car
```

```
##
## Attaching package: 'mice'
```

```
## The following object is masked from 'package:tidyr':
##
##     complete
```

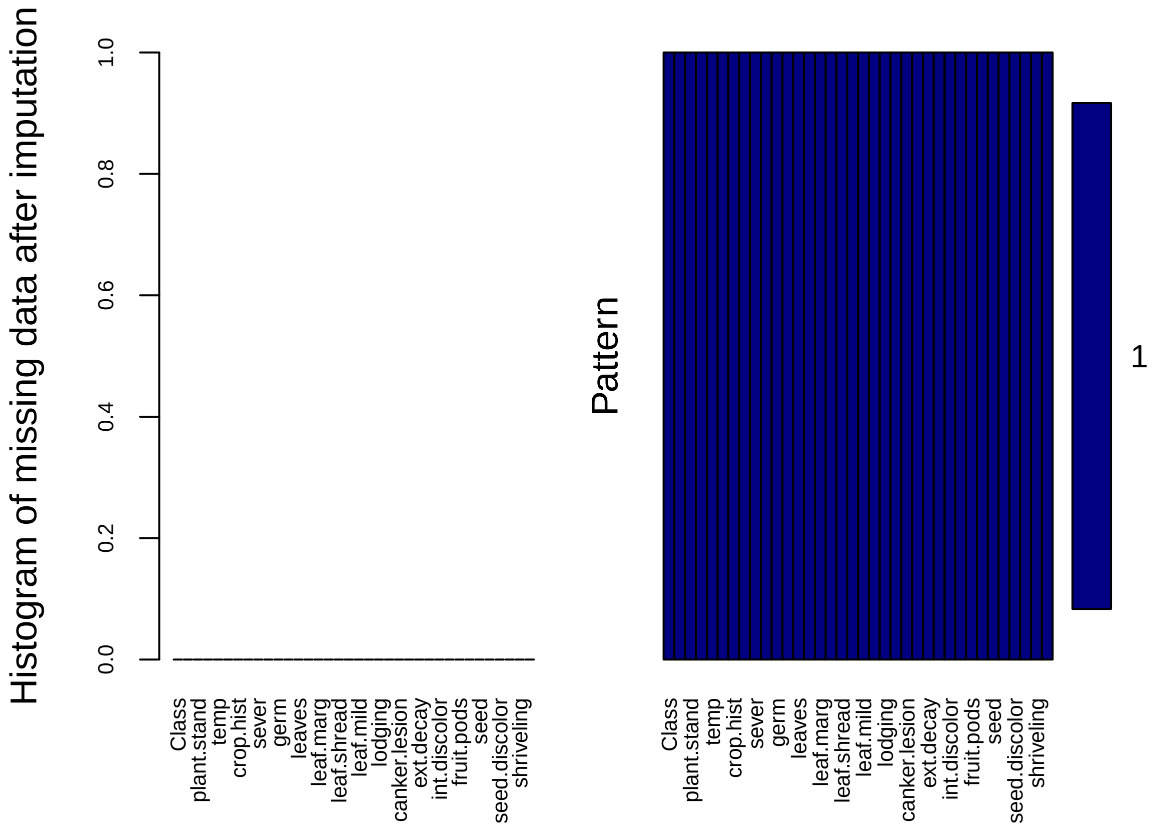
```
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
comp.data <- mice(Soybean,m=2,maxit=10,meth='pmm',seed=500)
```

```
##
##  iter imp variable
##   1   1  date  plant.stand  precip*  temp*  hail*  crop.hist*  area.dam  sever*  seed.tmt*  germ*  plant.growth*  leaf.halo*  leaf.marg
##   1   2  date  plant.stand*  precip  temp*  hail*  crop.hist*  area.dam  sever*  seed.tmt*  germ*  plant.growth*  leaf.halo  leaf.marg
##   2   1  date*  plant.stand*  precip*  temp*  hail*  crop.hist*  area.dam*  sever*  seed.tmt*  germ*  plant.growth*  leaf.halo*  leaf.m
##   2   2  date*  plant.stand*  precip*  temp*  hail*  crop.hist*  area.dam*  sever*  seed.tmt*  germ*  plant.growth*  leaf.halo*  leaf.m
##   3   1  date*  plant.stand*  precip*  temp*  hail*  crop.hist*  area.dam*  sever*  seed.tmt*  germ*  plant.growth*  leaf.halo*  leaf.m
##   3   2  date*  plant.stand*  precip*  temp*  hail*  crop.hist*  area.dam*  sever*  seed.tmt*  germ*  plant.growth*  leaf.halo*  leaf.m
##   4   1  date*  plant.stand*  precip*  temp*  hail*  crop.hist*  area.dam*  sever*  seed.tmt*  germ*  plant.growth*  leaf.halo*  leaf.m
##   4   2  date*  plant.stand*  precip*  temp*  hail*  crop.hist*  area.dam*  sever*  seed.tmt*  germ*  plant.growth*  leaf.halo*  leaf.m
##   5   1  date*  plant.stand*  precip*  temp*  hail*  crop.hist*  area.dam*  sever*  seed.tmt*  germ*  plant.growth*  leaf.halo*  leaf.m
##   5   2  date*  plant.stand*  precip*  temp*  hail*  crop.hist*  area.dam*  sever*  seed.tmt*  germ*  plant.growth*  leaf.halo*  leaf.m
##   6   1  date*  plant.stand*  precip*  temp*  hail*  crop.hist*  area.dam*  sever*  seed.tmt*  germ*  plant.growth*  leaf.halo*  leaf.m
##   6   2  date*  plant.stand*  precip*  temp*  hail*  crop.hist*  area.dam*  sever*  seed.tmt*  germ*  plant.growth*  leaf.halo*  leaf.m
##   7   1  date*  plant.stand*  precip*  temp*  hail*  crop.hist*  area.dam*  sever*  seed.tmt*  germ*  plant.growth*  leaf.halo*  leaf.m
##   7   2  date*  plant.stand*  precip*  temp*  hail*  crop.hist*  area.dam*  sever*  seed.tmt*  germ*  plant.growth*  leaf.halo*  leaf.m
##   8   1  date*  plant.stand*  precip*  temp*  hail*  crop.hist*  area.dam*  sever*  seed.tmt*  germ*  plant.growth*  leaf.halo*  leaf.m
##   8   2  date*  plant.stand*  precip*  temp*  hail*  crop.hist*  area.dam*  sever*  seed.tmt*  germ*  plant.growth*  leaf.halo*  leaf.m
##   9   1  date*  plant.stand*  precip*  temp*  hail*  crop.hist*  area.dam*  sever*  seed.tmt*  germ*  plant.growth*  leaf.halo*  leaf.m
##   9   2  date*  plant.stand*  precip*  temp*  hail*  crop.hist*  area.dam*  sever*  seed.tmt*  germ*  plant.growth*  leaf.halo*  leaf.m
##   10   1  date*  plant.stand*  precip*  temp*  hail*  crop.hist*  area.dam*  sever*  seed.tmt*  germ*  plant.growth*  leaf.halo*  leaf.
##   10   2  date*  plant.stand*  precip*  temp*  hail*  crop.hist*  area.dam*  sever*  seed.tmt*  germ*  plant.growth*  leaf.halo*  leaf.
##  * Please inspect the loggedEvents
```

```
Soybean.imputed = complete(comp.data)
```

```
##### Let's check if there is still any missing data, using VIM::aggr
#library(VIM)
ggr_plot <- aggr(Soybean.imputed,
                 col=c('navyblue','red'),
                 numbers=TRUE,
                 sortVars=TRUE,
                 labels=names(Soybean),
                 cex.axis=.7, gap=3,
                 ylab=c("Histogram of missing data after imputation","Pattern"))
```



Any missing data?

```
##
##  Variables sorted by number of missings:
##         Variable Count
##            Class     0
##             date     0
##      plant.stand     0
##           precip     0
##             temp     0
##             hail     0
##        crop.hist     0
##         area.dam     0
##            sever     0
##         seed.tmt     0
##             germ     0
##     plant.growth     0
##           leaves     0
##        leaf.halo     0
##        leaf.marg     0
##        leaf.size     0
##      leaf.shread     0
##        leaf.malf     0
##        leaf.mild     0
##             stem     0
##          lodging     0
##     stem.cankers     0
##    canker.lesion     0
##  fruiting.bodies     0
##        ext.decay     0
##         mycelium     0
##     int.discolor     0
##        sclerotia     0
##       fruit.pods     0
##      fruit.spots     0
##             seed     0
##      mold.growth     0
##    seed.discolor     0
##        seed.size     0
##       shriveling     0
##            roots     0
```

There is no missing data – all the NAs have been assigned imputed values.