DATA624 Homework 3

Bin Lin

2018-3-25

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:

```
#install.packages("mlbench")
#install.packages("kableExtra")
library(mlbench)
library(kableExtra)
library(knitr)

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 ...
```

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

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

Approaches: In order to understand the distribution of the predictor variable, the most straight forward method is to use the histogram obtain the frequencies of each predictor variables. The boxplot and summary statistics are useful in terms of searching for outliers and quantiles.And the correlation table tells us the relationship between each variables.

Interpretation: From the histogram, it shows the Fe and Ba variable contains lots of zeros, which make their graphs highly skewed to the right (K too). Most of the variables including RI, NA, AI, SI, CA have peaks in the center of the distribution. They appear to be more normally distributed. One exception is Mg, which has a trough in the center, but peaks on both ends. The summary statistics and boxplot have the same pattern as what is shown on the histogram, in addition to that, is also tell us there is lots of outliers in variable Ri, Al, Ca, Ba, Fe. The correlation table tell us that most of the variables are not related to each other, except the pair between RI and Ca, the correlation coefficient appear to be 0.7, which is moderately strong.

```r
glass <- subset(Glass, select = -Type)
predictors <- colnames(glass)

par(mfrow = c(3, 3))
for(i in 1:9)
  {
  hist(glass[,i], main = predictors[i])
}
```
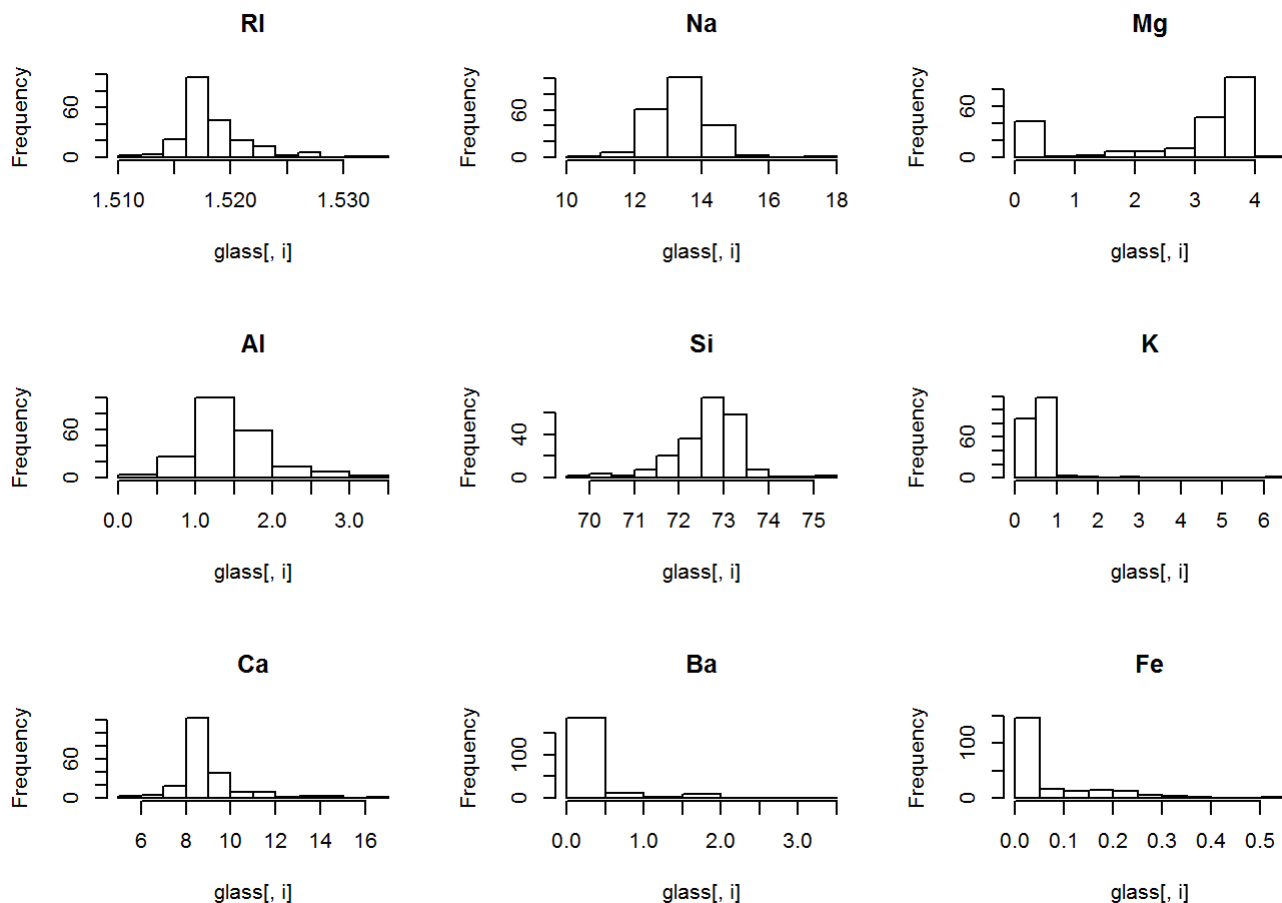


```r
par(mfcol=c(3, 3))
for (i in 1:9)
  {boxplot(glass[,i], main = predictors[i])}
```
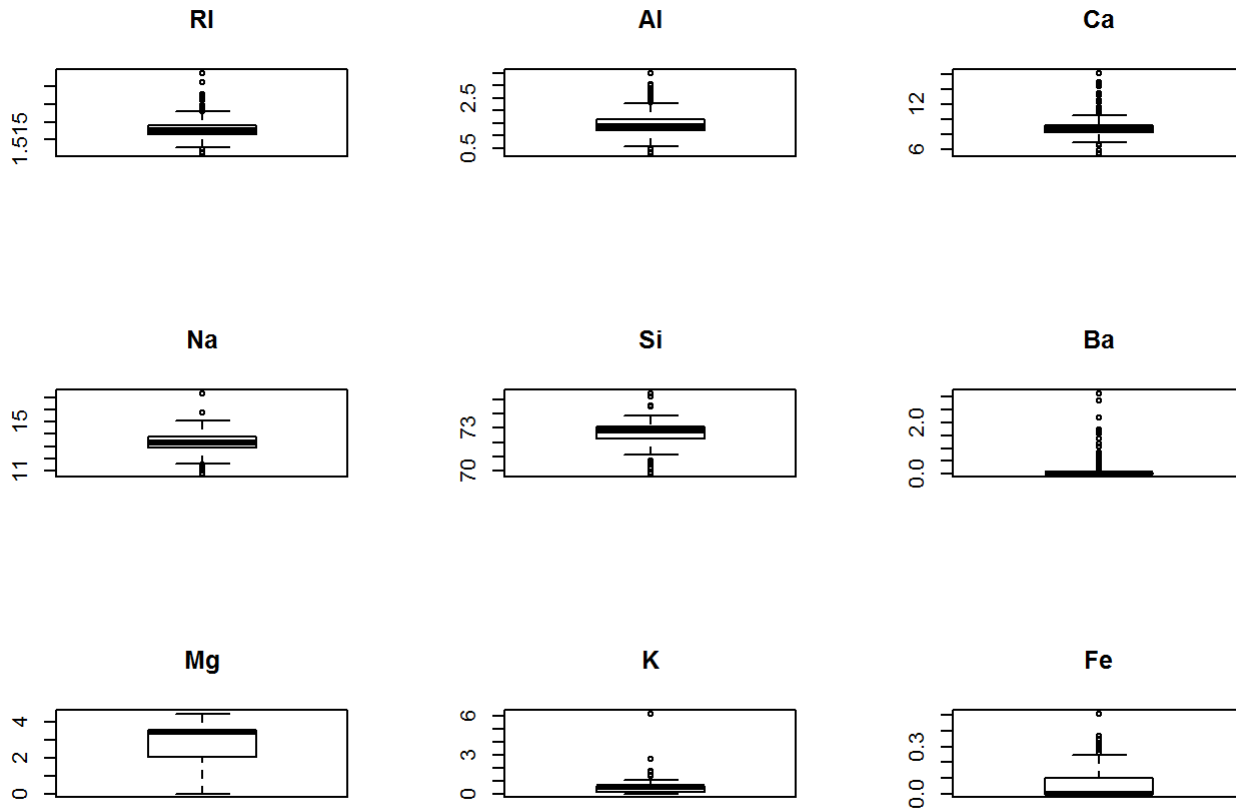
```
kable(summary(glass))
```

RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
Min. :1.511	Min. :10.73	Min. :0.000	Min. :0.290	Min. :69.81	Min. :0.0000	Min. : 5.430	Min. :0.000	Min. :0.00000
1st Qu.:1.517	1st Qu.:12.91	1st Qu.:2.115	1st Qu.:1.190	1st Qu.:72.28	1st Qu.:0.1225	1st Qu.: 8.240	1st Qu.:0.000	1st Qu.:0.00000
Median :1.518	Median :13.30	Median :3.480	Median :1.360	Median :72.79	Median :0.5550	Median : 8.600	Median :0.000	Median :0.00000
Mean :1.518	Mean :13.41	Mean :2.685	Mean :1.445	Mean :72.65	Mean :0.4971	Mean : 8.957	Mean :0.175	Mean :0.05701
3rd Qu.:1.519	3rd Qu.:13.82	3rd Qu.:3.600	3rd Qu.:1.630	3rd Qu.:73.09	3rd Qu.:0.6100	3rd Qu.: 9.172	3rd Qu.:0.000	3rd Qu.:0.10000
Max. :1.534	Max. :17.38	Max. :4.490	Max. :3.500	Max. :75.41	Max. :6.2100	Max. :16.190	Max. :3.150	Max. :0.51000

```
cor.table <- cor(glass, use = "pairwise", method = "spearman")
kable(round(cor.table, 2))
```

	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
RI	1.00	0.03	0.14	-0.49	-0.53	-0.29	0.70	-0.18	0.10
Na	0.03	1.00	-0.13	0.14	-0.27	-0.58	0.03	0.41	-0.22
Mg	0.14	-0.13	1.00	-0.51	-0.34	0.20	-0.29	-0.46	0.10
Al	-0.49	0.14	-0.51	1.00	0.20	0.15	-0.28	0.47	-0.08
Si	-0.53	-0.27	-0.34	0.20	1.00	0.00	-0.22	0.17	-0.07
K	-0.29	-0.58	0.20	0.15	0.00	1.00	-0.47	-0.26	0.09
Ca	0.70	0.03	-0.29	-0.28	-0.22	-0.47	1.00	-0.01	0.11
Ba	-0.18	0.41	-0.46	0.47	0.17	-0.26	-0.01	1.00	0.01
Fe	0.10	-0.22	0.10	-0.08	-0.07	0.09	0.11	0.01	1.00

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

Box-Cox transformation can be used to normalize the data, which are highly skewed. While Spatial Sign Transformation can be used to lower the effects of outliers, if the model is considered to be sensitive to outliers.

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)
library(ggplot2)
library(lattice)
data(Soybean)
str(Soybean)
```

```
## 'data.frame':    683 obs. of  36 variables:
## $ Class          : Factor w/ 19 levels "2-4-d-injury",..: 11 11 11 11 11 11 11 11 11 11 ...
## $ date           : Factor w/ 7 levels "0","1","2","3",..: 7 5 4 4 7 6 6 5 7 5 ...
## $ plant.stand    : Ord.factor w/ 2 levels "0"<"1": 1 1 1 1 1 1 1 1 1 1 ...
## $ precip         : Ord.factor w/ 3 levels "0"<"1"<"2": 3 3 3 3 3 3 3 3 3 3 ...
## $ temp           : Ord.factor w/ 3 levels "0"<"1"<"2": 2 2 2 2 2 2 2 2 2 2 ...
## $ hail           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 1 1 ...
## $ crop.hist      : Factor w/ 4 levels "0","1","2","3": 2 3 2 2 3 4 3 2 4 3 ...
## $ area.dam       : Factor w/ 4 levels "0","1","2","3": 2 1 1 1 1 1 1 1 1 1 ...
## $ sever          : Factor w/ 3 levels "0","1","2": 2 3 3 3 2 2 2 2 2 3 ...
## $ seed.tmt       : Factor w/ 3 levels "0","1","2": 1 2 2 1 1 1 2 1 2 1 ...
## $ germ           : Ord.factor w/ 3 levels "0"<"1"<"2": 1 2 3 2 3 2 1 3 2 3 ...
## $ plant.growth   : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
## $ leaves         : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
## $ leaf.halo      : Factor w/ 3 levels "0","1","2": 1 1 1 1 1 1 1 1 1 1 ...
## $ leaf.marg      : Factor w/ 3 levels "0","1","2": 3 3 3 3 3 3 3 3 3 3 ...
## $ leaf.size      : Ord.factor w/ 3 levels "0"<"1"<"2": 3 3 3 3 3 3 3 3 3 3 ...
## $ leaf.shread    : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
## $ leaf.malf      : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
## $ leaf.mild      : Factor w/ 3 levels "0","1","2": 1 1 1 1 1 1 1 1 1 1 ...
## $ stem           : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
## $ lodging        : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 2 1 1 1 ...
## $ stem.cankers   : Factor w/ 4 levels "0","1","2","3": 4 4 4 4 4 4 4 4 4 4 ...
## $ canker.lesion  : Factor w/ 4 levels "0","1","2","3": 2 2 1 1 2 1 2 2 2 2 ...
## $ fruiting.bodies: Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
## $ ext.decay      : Factor w/ 3 levels "0","1","2": 2 2 2 2 2 2 2 2 2 2 ...
## $ mycelium       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
## $ int.discolor   : Factor w/ 3 levels "0","1","2": 1 1 1 1 1 1 1 1 1 1 ...
## $ sclerotia      : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
## $ fruit.pods     : Factor w/ 4 levels "0","1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
## $ fruit.spots    : Factor w/ 4 levels "0","1","2","4": 4 4 4 4 4 4 4 4 4 4 ...
## $ seed           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
## $ mold.growth    : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
## $ seed.discolor  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
## $ seed.size      : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
## $ shriveling     : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
## $ roots          : Factor w/ 3 levels "0","1","2": 1 1 1 1 1 1 1 1 1 1 ...
```

```
head(Soybean, 30)
```

```
##                      Class date plant.stand precip temp hail crop.hist
## 1  diaporthe-stem-canker  6           0      2    1    0         1
## 2  diaporthe-stem-canker  4           0      2    1    0         2
## 3  diaporthe-stem-canker  3           0      2    1    0         1
## 4  diaporthe-stem-canker  3           0      2    1    0         1
## 5  diaporthe-stem-canker  6           0      2    1    0         2
## 6  diaporthe-stem-canker  5           0      2    1    0         3
## 7  diaporthe-stem-canker  5           0      2    1    0         2
## 8  diaporthe-stem-canker  4           0      2    1    1         1
## 9  diaporthe-stem-canker  6           0      2    1    0         3
## 10 diaporthe-stem-canker  4           0      2    1    0         2
## 11          charcoal-rot  6           0      0    2    0         1
## 12          charcoal-rot  4           0      0    1    1         1
## 13          charcoal-rot  3           0      0    1    0         1
## 14          charcoal-rot  6           0      0    1    1         3
## 15          charcoal-rot  6           0      0    2    0         1
## 16          charcoal-rot  5           0      0    2    1         3
## 17          charcoal-rot  6           0      0    2    1         0
## 18          charcoal-rot  4           0      0    1    0         2
## 19          charcoal-rot  3           0      0    2    0         2
## 20          charcoal-rot  5           0      0    2    1         2
## 21  rhizoctonia-root-rot  1           1      2    0    0         2
## 22  rhizoctonia-root-rot  1           1      2    0    0         1
## 23  rhizoctonia-root-rot  3           0      2    0    1         3
## 24  rhizoctonia-root-rot  0           1      2    0    0         0
## 25  rhizoctonia-root-rot  0           1      2    0    0         1
## 26  rhizoctonia-root-rot  1           1      2    0    0         3
## 27  rhizoctonia-root-rot  1           1      2    0    0         0
## 28  rhizoctonia-root-rot  2           1      2    0    0         2
## 29  rhizoctonia-root-rot  1           1      2    0    0         1
## 30  rhizoctonia-root-rot  2           1      2    0    0         1
##    area.dam sever seed.tmt germ plant.growth leaves leaf.halo leaf.marg
## 1         1     1        0    0            1      1         0         2
## 2         0     2        1    1            1      1         0         2
## 3         0     2        1    2            1      1         0         2
## 4         0     2        0    1            1      1         0         2
## 5         0     1        0    2            1      1         0         2
## 6         0     1        0    1            1      1         0         2
## 7         0     1        1    0            1      1         0         2
## 8         0     1        0    2            1      1         0         2
## 9         0     1        1    1            1      1         0         2
## 10        0     2        0    2            1      1         0         2
## 11        3     1        1    0            1      1         0         2
## 12        3     1        1    1            1      1         0         2
## 13        2     1        0    0            1      1         0         2
## 14        3     1        1    0            1      1         0         2
## 15        3     1        1    1            1      1         0         2
## 16        3     1        1    2            1      1         0         2
## 17        2     1        0    0            1      1         0         2
## 18        2     1        0    1            1      1         0         2
## 19        2     1        0    2            1      1         0         2
## 20        2     1        0    2            1      1         0         2
## 21        1     2        0    2            1      0         0         2
```

3/27/2018   DATA624 Homework 3

```
## 22        1     2       0     1          1      0         0         2
## 23        1     2       0     1          1      0         0         2
## 24        1     1       1     2          1      0         0         2
## 25        1     2       1     2          1      0         0         2
## 26        1     2       0     2          1      0         0         2
## 27        1     1       0     1          1      0         0         2
## 28        1     1       0     1          1      0         0         2
## 29        1     2       0     2          1      0         0         2
## 30        1     2       0     2          1      0         0         2
##      leaf.size leaf.shread leaf.malf leaf.mild stem lodging stem.cankers
## 1          2           0         0         0    1       1            3
## 2          2           0         0         0    1       0            3
## 3          2           0         0         0    1       0            3
## 4          2           0         0         0    1       0            3
## 5          2           0         0         0    1       0            3
## 6          2           0         0         0    1       0            3
## 7          2           0         0         0    1       1            3
## 8          2           0         0         0    1       0            3
## 9          2           0         0         0    1       0            3
## 10         2           0         0         0    1       0            3
## 11         2           0         0         0    1       0            0
## 12         2           0         0         0    1       1            0
## 13         2           0         0         0    1       0            0
## 14         2           0         0         0    1       0            0
## 15         2           0         0         0    1       0            0
## 16         2           0         0         0    1       0            0
## 17         2           0         0         0    1       1            0
## 18         2           0         0         0    1       0            0
## 19         2           0         0         0    1       0            0
## 20         2           0         0         0    1       0            0
## 21         2           0         0         0    1       0            1
## 22         2           0         0         0    1       0            1
## 23         2           0         0         0    1       1            1
## 24         2           0         0         0    1       0            1
## 25         2           0         0         0    1       0            1
## 26         2           0         0         0    1       0            1
## 27         2           0         0         0    1       0            1
## 28         2           0         0         0    1       0            1
## 29         2           0         0         0    1       0            1
## 30         2           0         0         0    1       0            1
##      canker.lesion fruiting.bodies ext.decay mycelium int.discolor sclerotia
## 1                1               1         1        0            0         0
## 2                1               1         1        0            0         0
## 3                0               1         1        0            0         0
## 4                0               1         1        0            0         0
## 5                1               1         1        0            0         0
## 6                0               1         1        0            0         0
## 7                1               1         1        0            0         0
## 8                1               1         1        0            0         0
## 9                1               1         1        0            0         0
## 10               1               1         1        0            0         0
## 11               3               0         0        0            2         1
## 12               3               0         0        0            2         1
## 13               3               0         0        0            2         1
```

```
## 14             3            0       0       0         2       1
## 15             3            0       0       0         2       1
## 16             3            0       0       0         2       1
## 17             3            0       0       0         2       1
## 18             3            0       0       0         2       1
## 19             3            0       0       0         2       1
## 20             3            0       0       0         2       1
## 21             1            0       1       1         0       0
## 22             1            0       1       0         0       0
## 23             1            0       1       1         0       0
## 24             1            0       1       0         0       0
## 25             1            0       1       0         0       0
## 26             1            0       1       0         0       0
## 27             1            0       1       0         0       0
## 28             1            0       1       0         0       0
## 29             1            0       1       0         0       0
## 30             1            0       1       0         0       0
##     fruit.pods fruit.spots seed mold.growth seed.discolor seed.size
## 1            0            4    0           0             0         0
## 2            0            4    0           0             0         0
## 3            0            4    0           0             0         0
## 4            0            4    0           0             0         0
## 5            0            4    0           0             0         0
## 6            0            4    0           0             0         0
## 7            0            4    0           0             0         0
## 8            0            4    0           0             0         0
## 9            0            4    0           0             0         0
## 10           0            4    0           0             0         0
## 11           0            4    0           0             0         0
## 12           0            4    0           0             0         0
## 13           0            4    0           0             0         0
## 14           0            4    0           0             0         0
## 15           0            4    0           0             0         0
## 16           0            4    0           0             0         0
## 17           0            4    0           0             0         0
## 18           0            4    0           0             0         0
## 19           0            4    0           0             0         0
## 20           0            4    0           0             0         0
## 21           3            4    0           0             0         0
## 22           3            4    0           0             0         0
## 23           3            4    0           0             0         0
## 24           3            4    0           0             0         0
## 25           3            4    0           0             0         0
## 26           3            4    0           0             0         0
## 27           3            4    0           0             0         0
## 28           3            4    0           0             0         0
## 29           3            4    0           0             0         0
## 30           3            4    0           0             0         0
##     shriveling roots
## 1            0     0
## 2            0     0
## 3            0     0
## 4            0     0
## 5            0     0
```

```
## 6               0     0
## 7               0     0
## 8               0     0
## 9               0     0
## 10              0     0
## 11              0     0
## 12              0     0
## 13              0     0
## 14              0     0
## 15              0     0
## 16              0     0
## 17              0     0
## 18              0     0
## 19              0     0
## 20              0     0
## 21              0     0
## 22              0     0
## 23              0     0
## 24              0     0
## 25              0     0
## 26              0     0
## 27              0     0
## 28              0     0
## 29              0     0
## 30              0     0
```

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

Approaches: Predictors with degenerate distributions are those predictors whose variances tend to approach zero. . The fraction of unique values over the sample size is low (say 10%). . 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 caret package function nearZeroVar will return the column numbers of any predictors that fulfill the conditions the conditions above.

Interpretation: The following code shows that three variables leaf.mild, mycelium and sclerotia are predictors with degenerate distribution.

```
#install.packages('caret', repos='http://cran.rstudio.com/')
#install.packages("dplyr")
#install.packages("tidyr")
library(tidyr)
library(dplyr)
```

```
##
## 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
```

```
library(caret)

index <- nearZeroVar(Soybean)
colnames(Soybean)[index]
```

```
## [1] "leaf.mild" "mycelium"  "sclerotia"
```

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?

Approaches: I am using the dplyr package to handle the data tidying and transformation. In this package there are many methods which I can use. for example: summarize_all, mutate, arrange et cetera.

Interpretation: According to first list, we can tell that hail, sever, seed.tmt, and lodge are teh variables that miss lots of data, While Class and leaves do not have missing data at all. Based on second graph, we can tell that the patten of missing data is also related to the classes. phytophthora-rot has the most, followed by 2-4-d-injury , nematode, diaporthe-pod-&-stem-blight, herbicide-injury. Everything else have complete sets of data.

```
missing_values <- Soybean %>%
  select(everything()) %>%
  summarize_all(funs(sum(is.na(.))))

data.frame(sort(missing_values, decreasing = TRUE))
```

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

```
missing_classes <- Soybean %>%
  gather(key = predictors, value = value, -Class)%>%
  group_by(Class)%>%
  summarize(n = sum(is.na(value)))%>%
  mutate(n, missing = n/(nrow(Soybean) * 35))%>%
  arrange(desc(missing))
```

```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```

```
missing_classes
```

```
## # A tibble: 19 x 3
##    Class                        n missing
##    <fct>                    <int>    <dbl>
##  1 phytophthora-rot          1214   0.0508
##  2 2-4-d-injury               450   0.0188
##  3 cyst-nematode              336   0.0141
##  4 diaporthe-pod-&-stem-blight 177 0.00740
##  5 herbicide-injury           160  0.00669
##  6 alternarialeaf-spot          0   0.
##  7 anthracnose                  0   0.
##  8 bacterial-blight             0   0.
##  9 bacterial-pustule            0   0.
## 10 brown-spot                   0   0.
## 11 brown-stem-rot               0   0.
## 12 charcoal-rot                 0   0.
## 13 diaporthe-stem-canker        0   0.
## 14 downy-mildew                 0   0.
## 15 frog-eye-leaf-spot           0   0.
## 16 phyllosticta-leaf-spot       0   0.
## 17 powdery-mildew               0   0.
## 18 purple-seed-stain            0   0.
## 19 rhizoctonia-root-rot         0   0.
```

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

Approaches: We can use Mice package to impute the missing values. Mice means multivariate imputation by chained equations. To improve the running time, I only run one time of imputation and use the default function ppm. Ppm means predictive mean matching. " It is similar to the regression method except that for each missing value, it imputes a value randomly from a set of observed values whose predicted values are closest to the predicted value for the missing value from the simulated regression model" (Heitjan and Little 1991; Schenker and Taylor 1996).

Interpretation:

The Mice package is very useful in terms of imputating values. As the end result shows, there is no more missing value in the data frame.

```
library(mice)
```

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

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

```
#install.packages("mice")

Soybean_impute <- mice(Soybean, m=1, method = "pmm", print = F)
Soybean_impute <- complete(Soybean_impute)

result<- Soybean_impute %>%
  select(everything()) %>%
  summarize_all(funs(sum(is.na(.))))

data.frame(sort(result, decreasing = TRUE))
```

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