Data 624: Week 3 Homework

Angrand, Burke, Deboch, Groysman, Karr

October 12, 2019

Week 3 Assignment

Chapter 3 KJ 1 and 2

3.1 The UC Irvine Machine Learning Repository 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)
describe(Glass)
```

```
##       vars   n  mean   sd median trimmed  mad   min   max range  skew
## RI       1 214  1.52 0.00   1.52    1.52 0.00  1.51  1.53  0.02  1.60
## Na       2 214 13.41 0.82  13.30   13.38 0.64 10.73 17.38  6.65  0.45
## Mg       3 214  2.68 1.44   3.48    2.87 0.30  0.00  4.49  4.49 -1.14
## Al       4 214  1.44 0.50   1.36    1.41 0.31  0.29  3.50  3.21  0.89
## Si       5 214 72.65 0.77  72.79   72.71 0.57 69.81 75.41  5.60 -0.72
## K        6 214  0.50 0.65   0.56    0.43 0.17  0.00  6.21  6.21  6.46
## Ca       7 214  8.96 1.42   8.60    8.74 0.66  5.43 16.19 10.76  2.02
## Ba       8 214  0.18 0.50   0.00    0.03 0.00  0.00  3.15  3.15  3.37
## Fe       9 214  0.06 0.10   0.00    0.04 0.00  0.00  0.51  0.51  1.73
## Type*   10 214  2.54 1.71   2.00    2.31 1.48  1.00  6.00  5.00  1.04
##       kurtosis   se
## RI        4.72 0.00
## Na        2.90 0.06
## Mg       -0.45 0.10
## Al        1.94 0.03
## Si        2.82 0.05
## K        52.87 0.04
## Ca        6.41 0.10
## Ba       12.08 0.03
## Fe        2.52 0.01
## Type*    -0.29 0.12
```

```
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 ...
```

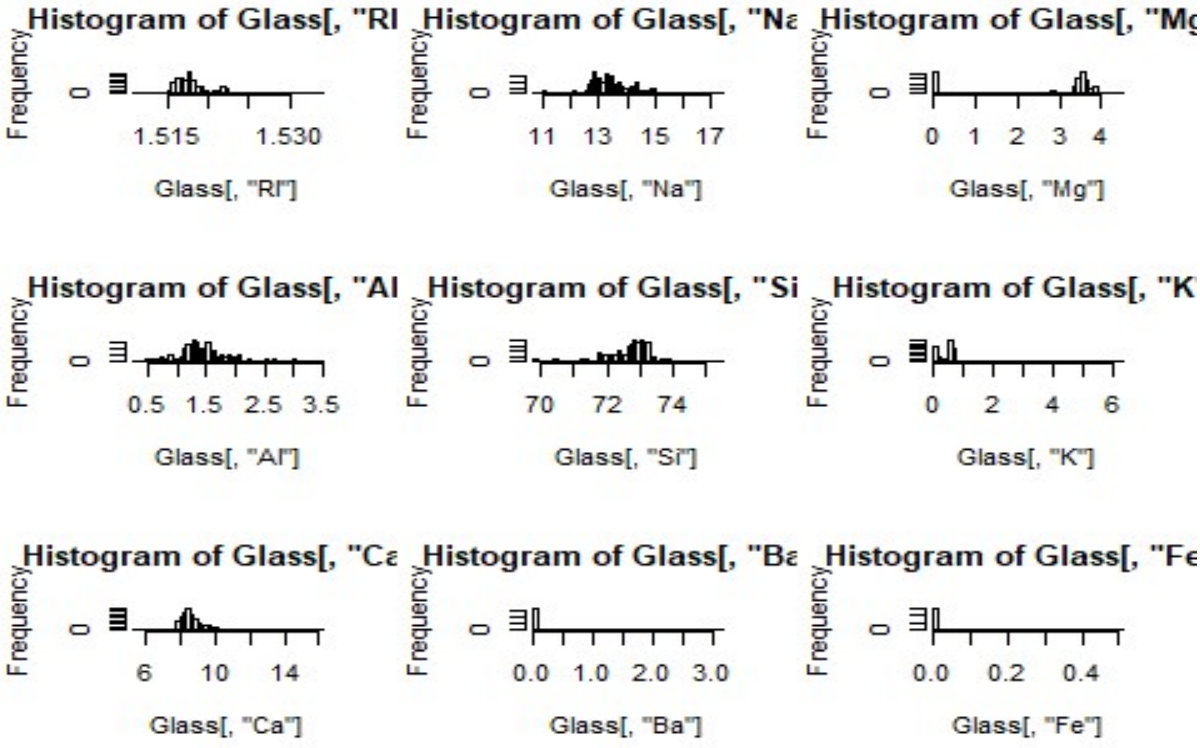
```
my_df <- data.frame(Glass[,1:9])

cor(my_df)
```

```
##                RI          Na           Mg          Al           Si
## RI  1.0000000000 -0.19188538 -0.122274039 -0.40732603 -0.54205220
## Na -0.1918853790  1.00000000 -0.273731961  0.15679367 -0.06980881
## Mg -0.1222740393 -0.27373196  1.000000000 -0.48179851 -0.16592672
## Al -0.4073260341  0.15679367 -0.481798509  1.00000000 -0.00552372
## Si -0.5420521997 -0.06980881 -0.165926723 -0.00552372  1.00000000
## K  -0.2898327111 -0.26608650  0.005395667  0.32595845 -0.19333085
## Ca  0.8104026963 -0.27544249 -0.443750026 -0.25959201 -0.20873215
## Ba -0.0003860189  0.32660288 -0.492262118  0.47940390 -0.10215131
## Fe  0.1430096093 -0.24134641  0.083059529 -0.07440215 -0.09420073
##                K          Ca            Ba           Fe
## RI -0.289832711  0.8104027 -0.0003860189  0.143009609
## Na -0.266086504 -0.2754425  0.3266028795 -0.241346411
## Mg  0.005395667 -0.4437500 -0.4922621178  0.083059529
## Al  0.325958446 -0.2595920  0.4794039017 -0.074402151
## Si -0.193330854 -0.2087322 -0.1021513105 -0.094200731
## K   1.000000000 -0.3178362 -0.0426180594 -0.007719049
## Ca -0.317836155  1.0000000 -0.1128409671  0.124968219
## Ba -0.042618059 -0.1128410  1.0000000000 -0.058691755
## Fe -0.007719049  0.1249682 -0.0586917554  1.000000000
```

- A data frame with 214 observation containing examples of the chemical analysis of 7 different types of glass.

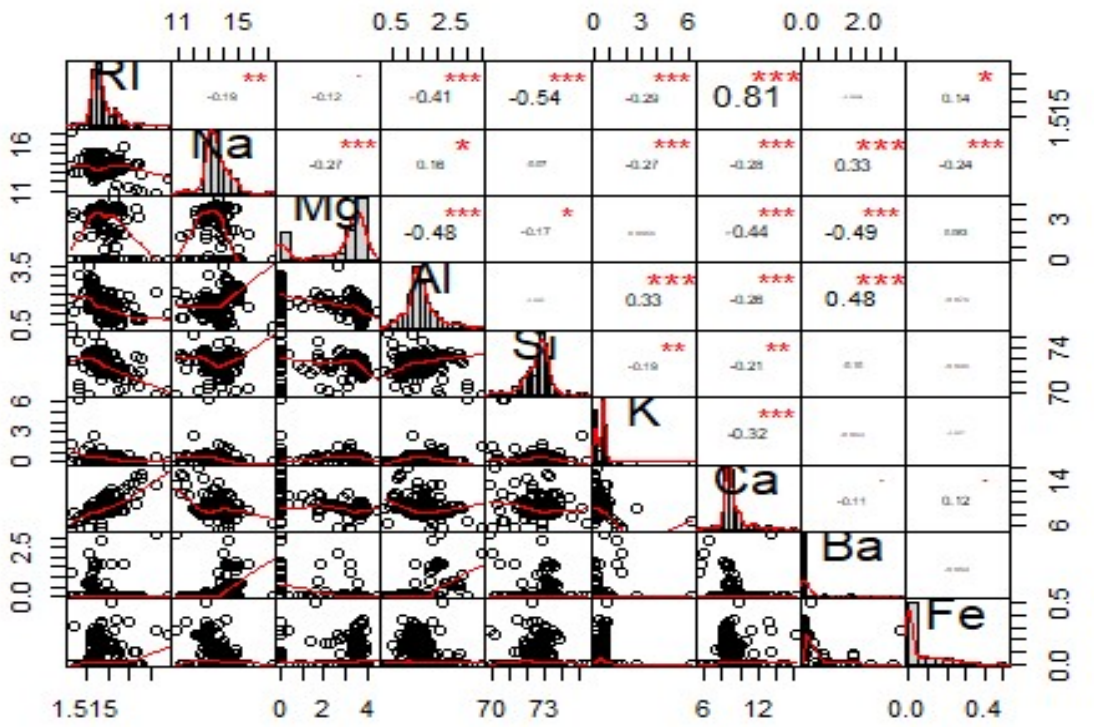
a. Using visualizations explore the predictor variables to understand their distributions as well as the relationships between predictors.

```
#histograms for each
#ZERO-INFLATED NEGATIVE BINOMIAL for Mg, Ba & Fe or is it a nuanced distribut
ion
par(mfrow = c(3,3))
hist(Glass[,'RI'],breaks=50)
hist(Glass[,'Na'],breaks=50)
hist(Glass[,'Mg'],breaks=50)
hist(Glass[,'Al'],breaks=50)
hist(Glass[,'Si'],breaks=50)
hist(Glass[,'K'],breaks=50)
hist(Glass[,'Ca'],breaks=50)
hist(Glass[,'Ba'],breaks=50)
hist(Glass[,'Fe'],breaks=50)
```



- There are a total of 214 glass samples taken with no instances of missing data for any of the predictor variables. Based upon their histograms and skewness, the predictors RI, Na, Al, Si & Ca display either either a normal distribution pattern or a distribution that could be transformed into a normal distribution pattern i.e. division by sqrt(s). The remaining predictor variables Mg, K, Ba & Fe display concentrations of 0 frequency.

```
my_df <- data.frame(Glass[,1:9])
chart.Correlation(my_df, histogram=TRUE, pch=19)
```



- From correlation we can see that:
  - RI is significantly positively correlated with CA and negatively correlated with AL,Si,K.
  - Na is Significantly positively correlated with Ba and negatively correlated with Mg,Al,K,Ca,Fe.
  - Mg is significantly negatively correlated with Ca,Ba ,Al.
  - Al is significantly positively correlated with K,Ba and negatively correlated with Ca.
  - Si is weakly negatively correlated with K and Cal.

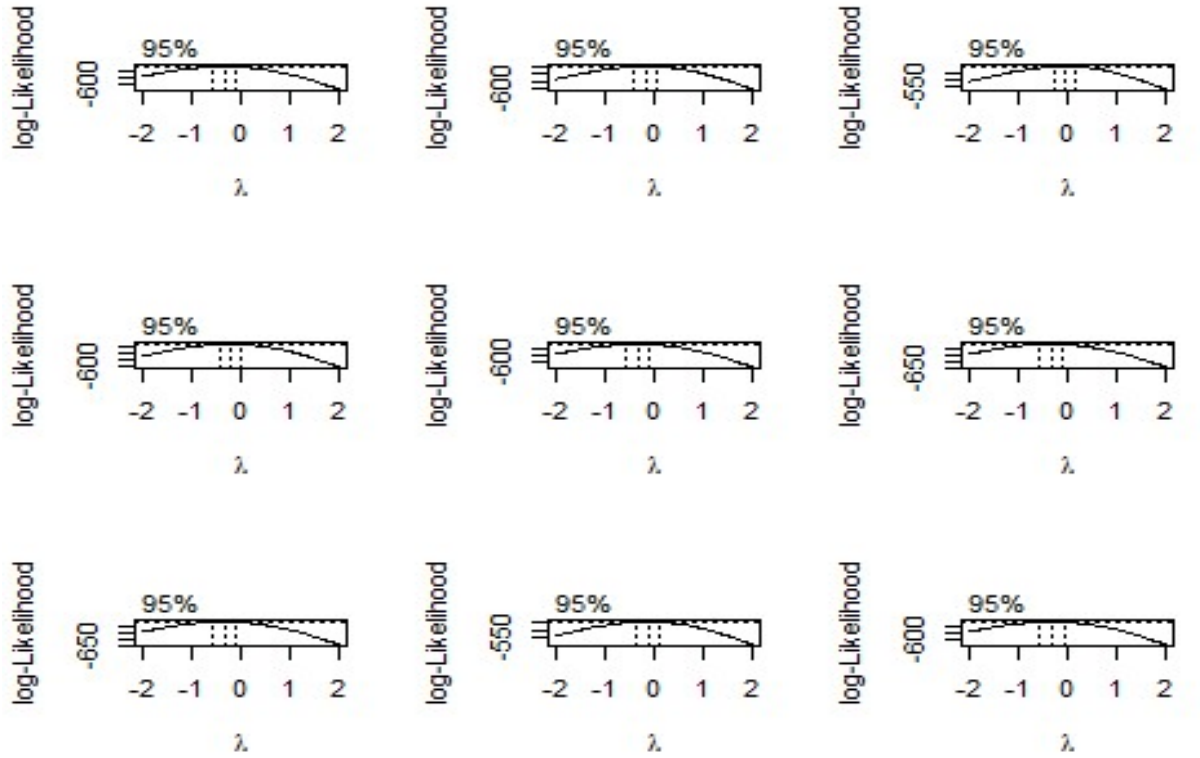
b. Do there appear to be any outliers in the data? Are any predictors skewed?

- From the above plot of histograms we can see that Mg,Si,K,Ca,Ba and Fe has outliers. Fe,Ba,Ca,K,Na,RI are positively skewed and Mg,Si are negatively skewed.

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

```
Glass$Type <- as.numeric(Glass$Type)
par(mfrow = c(3,3))
boxcox(Type~RI, data = Glass)
boxcox(Type~Na, data = Glass)
boxcox(Type~Mg, data = Glass)
boxcox(Type~Al, data = Glass)
```

```
boxcox(Type~Si, data = Glass)
boxcox(Type~K, data = Glass)
boxcox(Type~Ca, data = Glass)
boxcox(Type~Ba, data = Glass)
boxcox(Type~Fe, data = Glass)
```



- A better solution to handling the predictors with concentrations of 0 frequency is to use a zero-inflated binary distribution for continuous data. The two predictors with the greatest correlation are RI and Ca suggesting that in a multivariable regression model, one of these explanatory variables could be removed because it is strongly co-linear with the other thus having little to no loss of predictive ability to the model. Also, from the box cox transformation plot we can see that log transformation of Na, Mg and Ba will improve the model

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.

```
#Preliminary EDA
#Data Access
data(Soybean)
#Sampling
glimpse(Soybean)
```

```
## Observations: 683
## Variables: 36
## $ Class           <fct> diaporthe-stem-canker, diaporthe-stem-canker, ...
## $ date            <fct> 6, 4, 3, 3, 6, 5, 5, 4, 6, 4, 6, 4, 3, 6, 6, 5...
## $ plant.stand     <ord> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ precip          <ord> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0...
## $ temp            <ord> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 2, 2...
## $ hail            <fct> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 1...
## $ crop.hist       <fct> 1, 2, 1, 1, 2, 3, 2, 1, 3, 2, 1, 1, 1, 3, 1, 3...
## $ area.dam        <fct> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 3, 2, 3, 3, 3...
## $ sever           <fct> 1, 2, 2, 2, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1...
## $ seed.tmt        <fct> 0, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1...
## $ germ            <ord> 0, 1, 2, 1, 2, 1, 0, 2, 1, 2, 0, 1, 0, 0, 1, 2...
## $ plant.growth    <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ leaves          <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ leaf.halo       <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ leaf.marg       <fct> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2...
## $ leaf.size       <ord> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2...
## $ leaf.shread     <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ leaf.malf       <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ leaf.mild       <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ stem            <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ lodging         <fct> 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0...
## $ stem.cankers    <fct> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 0, 0, 0, 0, 0, 0...
## $ canker.lesion   <fct> 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 3, 3, 3, 3, 3, 3...
## $ fruiting.bodies <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0...
## $ ext.decay       <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0...
## $ mycelium        <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ int.discolor    <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 2, 2, 2, 2, 2...
## $ sclerotia       <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1...
## $ fruit.pods      <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ fruit.spots     <fct> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4...
## $ seed            <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ mold.growth     <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ seed.discolor   <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ seed.size       <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ shriveling      <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ roots           <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
```

```
#Shape
dim(Soybean)
```

```
## [1] 683  36
```

```
#Stats
describe(Soybean)
```

```
##                  vars   n mean   sd median trimmed  mad min max range
## Class*              1 683 9.30 5.51      8    9.18 7.41   1  19    18
## date*               2 682 4.55 1.69      5    4.62 1.48   1   7     6
## plant.stand*        3 647 1.45 0.50      1    1.44 0.00   1   2     1
```

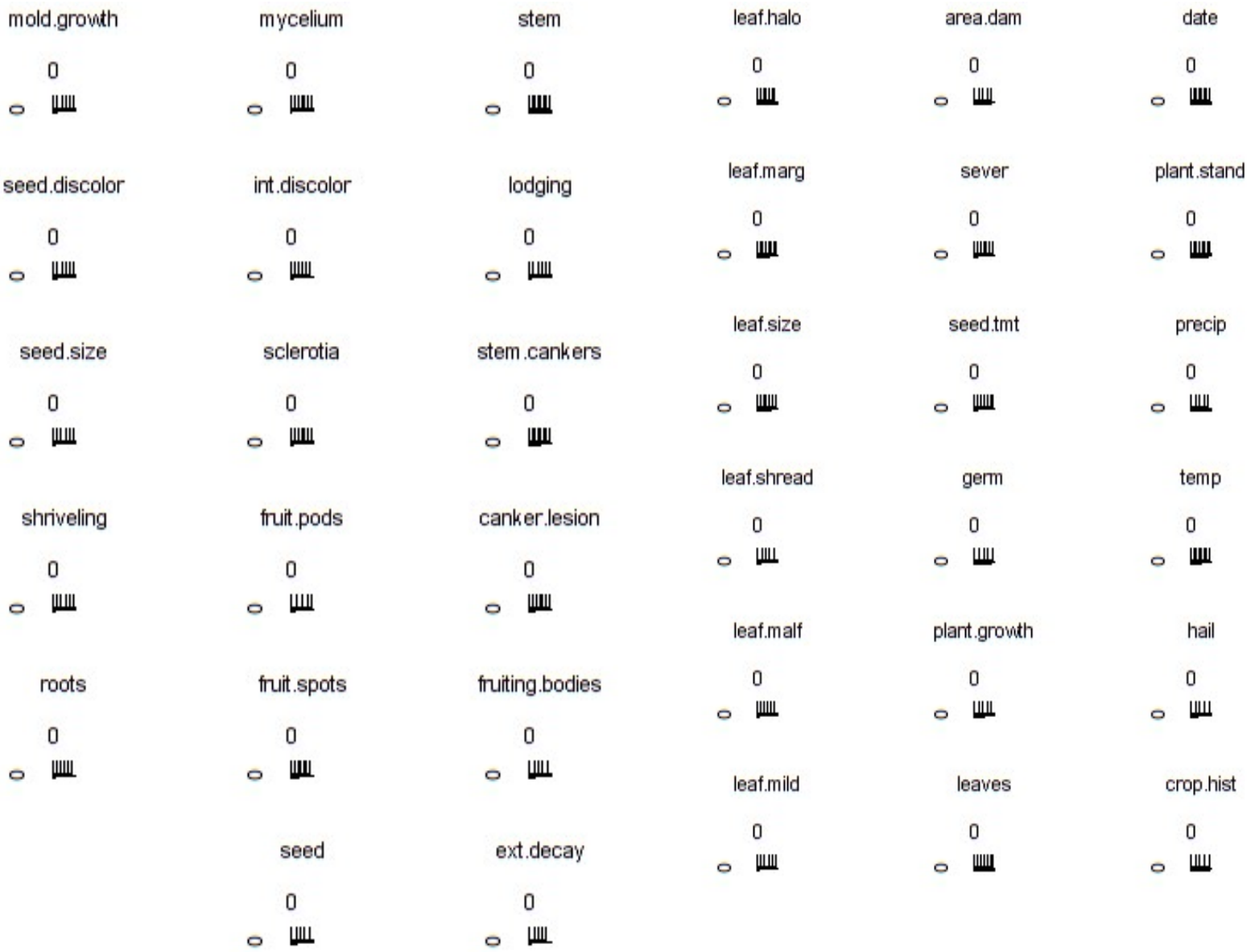
```
## precip*             4 645 2.60 0.69      3    2.74 0.00   1   3     2
## temp*               5 653 2.18 0.63      2    2.23 0.00   1   3     2
## hail*               6 562 1.23 0.42      1    1.16 0.00   1   2     1
## crop.hist*          7 667 2.88 0.98      3    2.98 1.48   1   4     3
## area.dam*           8 682 2.58 1.07      2    2.60 1.48   1   4     3
## sever*              9 562 1.73 0.60      2    1.69 0.00   1   3     2
## seed.tmt*          10 562 1.52 0.61      1    1.45 0.00   1   3     2
## germ*              11 571 2.05 0.79      2    2.06 1.48   1   3     2
## plant.growth*      12 667 1.34 0.47      1    1.30 0.00   1   2     1
## leaves*            13 683 1.89 0.32      2    1.98 0.00   1   2     1
## leaf.halo*         14 599 2.20 0.95      3    2.25 0.00   1   3     2
## leaf.marg*         15 599 1.77 0.96      1    1.72 0.00   1   3     2
## leaf.size*         16 599 2.28 0.61      2    2.34 0.00   1   3     2
## leaf.shread*       17 583 1.16 0.37      1    1.08 0.00   1   2     1
## leaf.malf*         18 599 1.08 0.26      1    1.00 0.00   1   2     1
## leaf.mild*         19 575 1.10 0.40      1    1.00 0.00   1   3     2
## stem*              20 667 1.56 0.50      2    1.57 0.00   1   2     1
## lodging*           21 562 1.07 0.26      1    1.00 0.00   1   2     1
## stem.cankers*      22 645 2.06 1.35      1    1.95 0.00   1   4     3
## canker.lesion*     23 645 1.98 1.08      2    1.85 1.48   1   4     3
## fruiting.bodies*   24 577 1.18 0.38      1    1.10 0.00   1   2     1
## ext.decay*         25 645 1.25 0.48      1    1.16 0.00   1   3     2
## mycelium*          26 645 1.01 0.10      1    1.00 0.00   1   2     1
## int.discolor*      27 645 1.13 0.42      1    1.00 0.00   1   3     2
## sclerotia*         28 645 1.03 0.17      1    1.00 0.00   1   2     1
## fruit.pods*        29 599 1.50 0.88      1    1.28 0.00   1   4     3
## fruit.spots*       30 577 1.85 1.17      1    1.69 0.00   1   4     3
## seed*              31 591 1.19 0.40      1    1.12 0.00   1   2     1
## mold.growth*       32 591 1.11 0.32      1    1.02 0.00   1   2     1
## seed.discolor*     33 577 1.11 0.31      1    1.02 0.00   1   2     1
## seed.size*         34 591 1.10 0.30      1    1.00 0.00   1   2     1
## shriveling*        35 577 1.07 0.25      1    1.00 0.00   1   2     1
## roots*             36 652 1.18 0.44      1    1.07 0.00   1   3     2
##                    skew kurtosis   se
## Class*             0.11    -1.38 0.21
## date*             -0.30    -0.90 0.06
## plant.stand*       0.19    -1.97 0.02
## precip*           -1.42     0.55 0.03
## temp*             -0.16    -0.58 0.02
## hail*              1.31    -0.29 0.02
## crop.hist*        -0.40    -0.92 0.04
## area.dam*          0.02    -1.29 0.04
## sever*             0.17    -0.56 0.03
## seed.tmt*          0.74    -0.44 0.03
## germ*             -0.09    -1.40 0.03
## plant.growth*      0.68    -1.54 0.02
## leaves*           -2.44     3.98 0.01
## leaf.halo*        -0.41    -1.76 0.04
## leaf.marg*         0.46    -1.75 0.04
## leaf.size*        -0.25    -0.63 0.02
```

```
## leaf.shread*      1.80     1.26 0.02
## leaf.malf*        3.22     8.35 0.01
## leaf.mild*        3.95    14.68 0.02
## stem*            -0.23    -1.95 0.02
## lodging*          3.23     8.42 0.01
## stem.cankers*     0.61    -1.51 0.05
## canker.lesion*    0.51    -1.24 0.04
## fruiting.bodies*  1.66     0.75 0.02
## ext.decay*        1.70     1.98 0.02
## mycelium*        10.20   102.18 0.00
## int.discolor*     3.34    10.57 0.02
## sclerotia*        5.40    27.19 0.01
## fruit.pods*       1.84     2.41 0.04
## fruit.spots*      0.95    -0.76 0.05
## seed*             1.54     0.37 0.02
## mold.growth*      2.43     3.93 0.01
## seed.discolor*    2.47     4.12 0.01
## seed.size*        2.66     5.10 0.01
## shriveling*       3.49    10.21 0.01
## roots*            2.46     5.49 0.02
```

- There are 19 classes, only the first 15 of which have been used in prior work. 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," etc.

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

```
df <- Soybean[,2:36]
par(mfrow = c(3, 6))
for (i in 1:ncol(df)) {
  barplot(table(df[,i]),ylab = names(df[i]))
}
```


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

nearZeroVar in R for the categorical variables

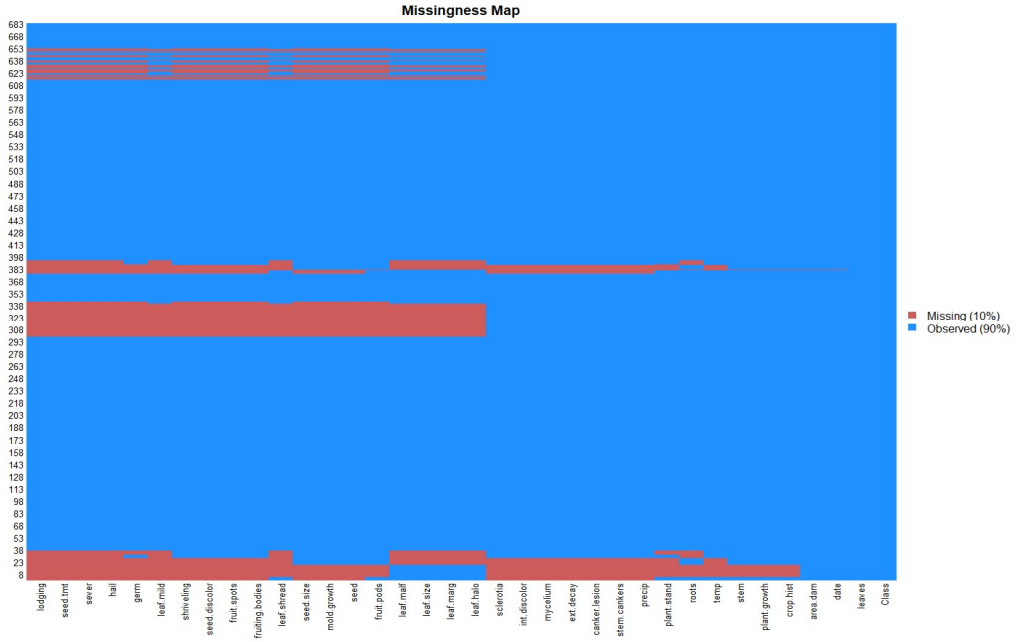
```
nearZeroVar(df,names = TRUE, saveMetrics=T)
```

```
##                   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
```

- There are few distributions degenerate . Specifically leaf.mild,mycelium and 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?

```
missmap(Soybean)
```




```
sort(colMeans(is.na(Soybean)),decreasing = T)
```

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

- Particularly hail, sever,seed.tmt,lodging, germ,leaf.mild fruiting.bodies, fruit.spots,seed.discolor,shriveling, leaf.shread, seed,mold.growth,seed.size,leaf.halo,are more likelly to be missing.

```
Soybean %>%
mutate(total = n()) %>%
group_by(Class) %>%
mutate(Missing = n(), Proportion=Missing/total) %>%
```

```
dplyr::select(Class, Missing, Proportion) %>%
unique() %>%
  arrange(-Proportion)
```

```
## # A tibble: 19 x 3
## # Groups:   Class [19]
##    Class                       Missing Proportion
##    <fct>                         <int>      <dbl>
##  1 brown-spot                       92     0.135
##  2 alternarialeaf-spot              91     0.133
##  3 frog-eye-leaf-spot               91     0.133
##  4 phytophthora-rot                 88     0.129
##  5 brown-stem-rot                   44     0.0644
##  6 anthracnose                      44     0.0644
##  7 diaporthe-stem-canker            20     0.0293
##  8 charcoal-rot                     20     0.0293
##  9 rhizoctonia-root-rot             20     0.0293
## 10 powdery-mildew                   20     0.0293
## 11 downy-mildew                     20     0.0293
## 12 bacterial-blight                 20     0.0293
## 13 bacterial-pustule                20     0.0293
## 14 purple-seed-stain                20     0.0293
## 15 phyllosticta-leaf-spot           20     0.0293
## 16 2-4-d-injury                     16     0.0234
## 17 diaporthe-pod-&-stem-blight      15     0.0220
## 18 cyst-nematode                    14     0.0205
## 19 herbicide-injury                  8     0.0117
```

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

- Drop the rows having missing values. After dropping, 562 observations remain.

```
Soybean_complete <- na.omit(Soybean)
head(Soybean_complete)
```

```
##                   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
##   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
##   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
##   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
##   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
##   shriveling roots
## 1          0     0
## 2          0     0
## 3          0     0
## 4          0     0
## 5          0     0
## 6          0     0
```

```
dim(Soybean_complete)
```

```
## [1] 562  36
```