Homework 7

Darian-Florian Voda

2022-12-09

Loading packages

```
library(dplyr)
library(ggplot2)
library(car)
library(tidyverse)
library(ggpubr)
library(rstatix)
```

Exercise 85: Outcome of lung ventilation

- Use the dataset „Discriminant-pulmonary.txt“
- Contains diverse parameters on patients receiving lung ventilation
  - Age
  - Oxygen concentration in blood
  - Body size
  - Aggressiveness of the ventilation
  - Ventilation time
- Not all patients survived. Can you predict, based on these parameters, whether a patient will survive or not?
- Also try standardization of variables!

Load packages

```
library(ggplot2)
library(klaR)
library(ggord)
library(psych)
library(MASS)
library(devtools)
library(tidyverse)
library(caret)
library(mosaic)
library(broom)
```

For this exercise, we will try to predict **using** scaled variables and **without** scaled variables.

Let's start without scaled variables

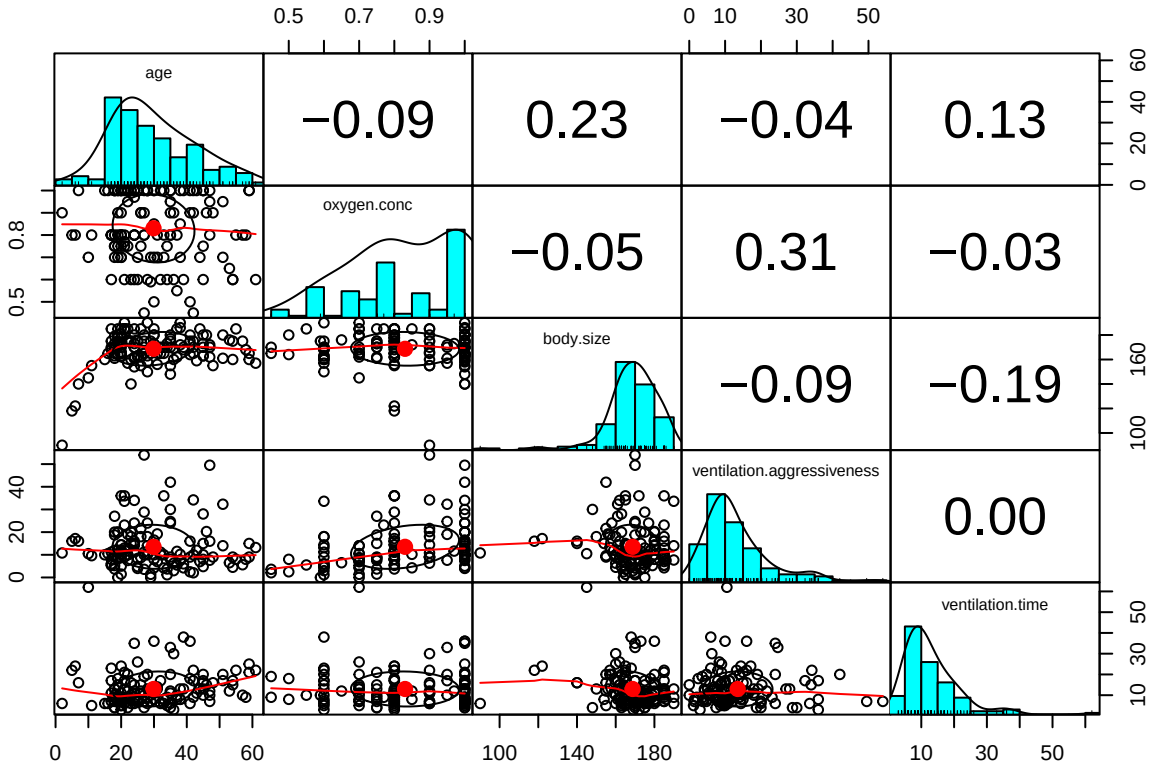
```
# Without standardization
data<-read.delim("C:/Users/daria/OneDrive/Desktop/Master - AppDS/Statistics/Datasets-20221007/Discrimina
                 stringsAsFactors=F)
# Look at the data
head(data)
```

```
##    outcome age oxygen.conc body.size ventilation.aggressiveness
## 1     dead  27        0.45       165                       3.60
## 2 survived  35        0.95       170                      24.70
## 3     dead  15        1.00       160                      10.00
## 4 survived  19        0.75       175                       8.25
## 5 survived  21        0.77       185                      23.10
## 6 survived  24        0.80       180                      14.40
##   ventilation.time
## 1               19
## 2               33
## 3                6
## 4                7
## 5                9
## 6                6
```

```
# Get type of the data
str(data)
```

```
## 'data.frame':    131 obs. of  6 variables:
##  $ outcome                   : chr  "dead" "survived" "dead" "survived" ...
##  $ age                       : int  27 35 15 19 21 24 21 24 20 27 ...
##  $ oxygen.conc               : num  0.45 0.95 1 0.75 0.77 0.8 0.75 0.8 1 0.8 ...
##  $ body.size                 : int  165 170 160 175 185 180 185 185 165 175 ...
##  $ ventilation.aggressiveness: num  3.6 24.7 10 8.25 23.1 14.4 9 16 5 28.8 ...
##  $ ventilation.time          : int  19 33 6 7 9 6 21 11 8 4 ...
```

```
pairs.panels(data[2:6],
             gap = 0,
             bg = c("red", "green")[data$outcome],
             pch = 21)
```




```
# 70% for training, 30% for testing
set.seed(123)
ind <- sample(2, nrow(data),
              replace = TRUE,
              prob = c(0.7, 0.3))
training <- data[ind==1,]
testing <- data[ind==2,]

linear <- lda(outcome~., training)
linear
```

```
## Call:
## lda(outcome ~ ., data = training)
##
## Prior probabilities of groups:
##     dead survived
##      0.5      0.5
##
## Group means:
##                age oxygen.conc body.size ventilation.aggressiveness
## dead     30.93617   0.8606383  166.2340                    15.07598
## survived 30.00000   0.8044681  172.1702                    10.06594
##          ventilation.time
## dead             15.31915
## survived         11.08511
##
```

```
## Coefficients of linear discriminants:
##                                     LD1
## age                          -0.008517239
## oxygen.conc                  -2.164301644
## body.size                     0.031798780
## ventilation.aggressiveness -0.054405556
## ventilation.time             -0.089326095
```

```
plot(linear)

paste("Prior probabilities of survival: 50%/50% rate of survival")
```

```
## [1] "Prior probabilities of survival: 50%/50% rate of survival"
```

```
# prediction
p <- predict(linear, training)
names(p)
```

```
## [1] "class"     "posterior" "x"
```

```
# Predicted classes
head(p$class, 6)
```

```
## [1] survived survived survived survived survived survived
## Levels: dead survived
```

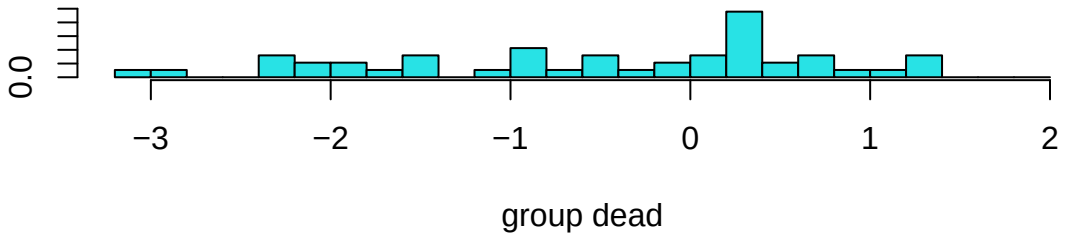
```
# Predicted probabilities
head(p$posterior, 6)
```

```
##         dead  survived
## 1  0.3379166 0.6620834
## 3  0.4373556 0.5626444
## 6  0.2726083 0.7273917
## 7  0.4374854 0.5625146
## 9  0.3881250 0.6118750
## 10 0.4460870 0.5539130
```

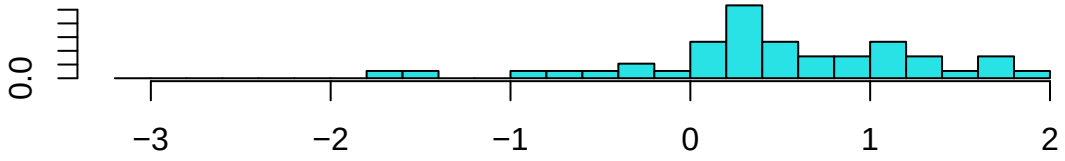
```
# Linear discriminants
head(p$x, 3)
```

```
##         LD1
## 1 0.6940451
## 3 0.2599359
## 6 1.0127322
```

```
# Stacked histogram for LD1
p <- predict(linear, training)
ldahist(data = p$x[,1], g = training$outcome)
```

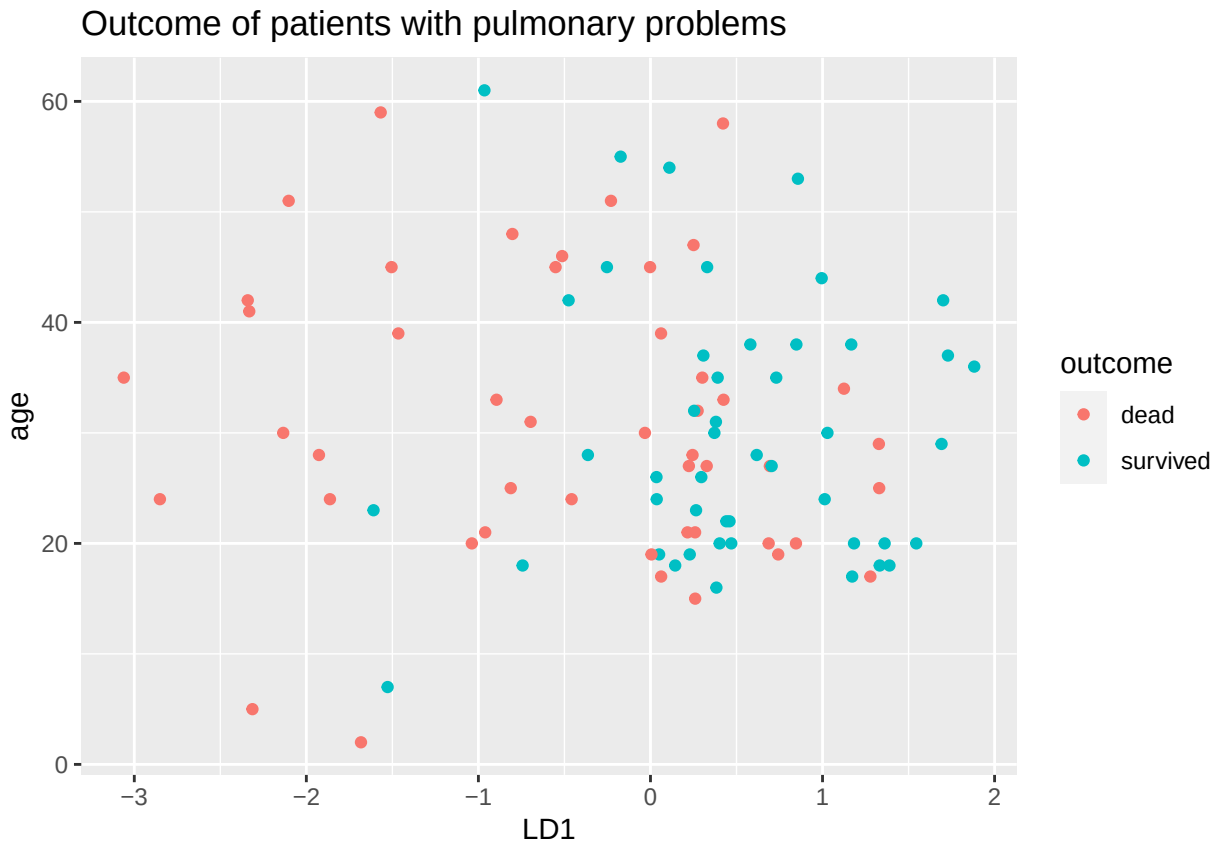


group dead

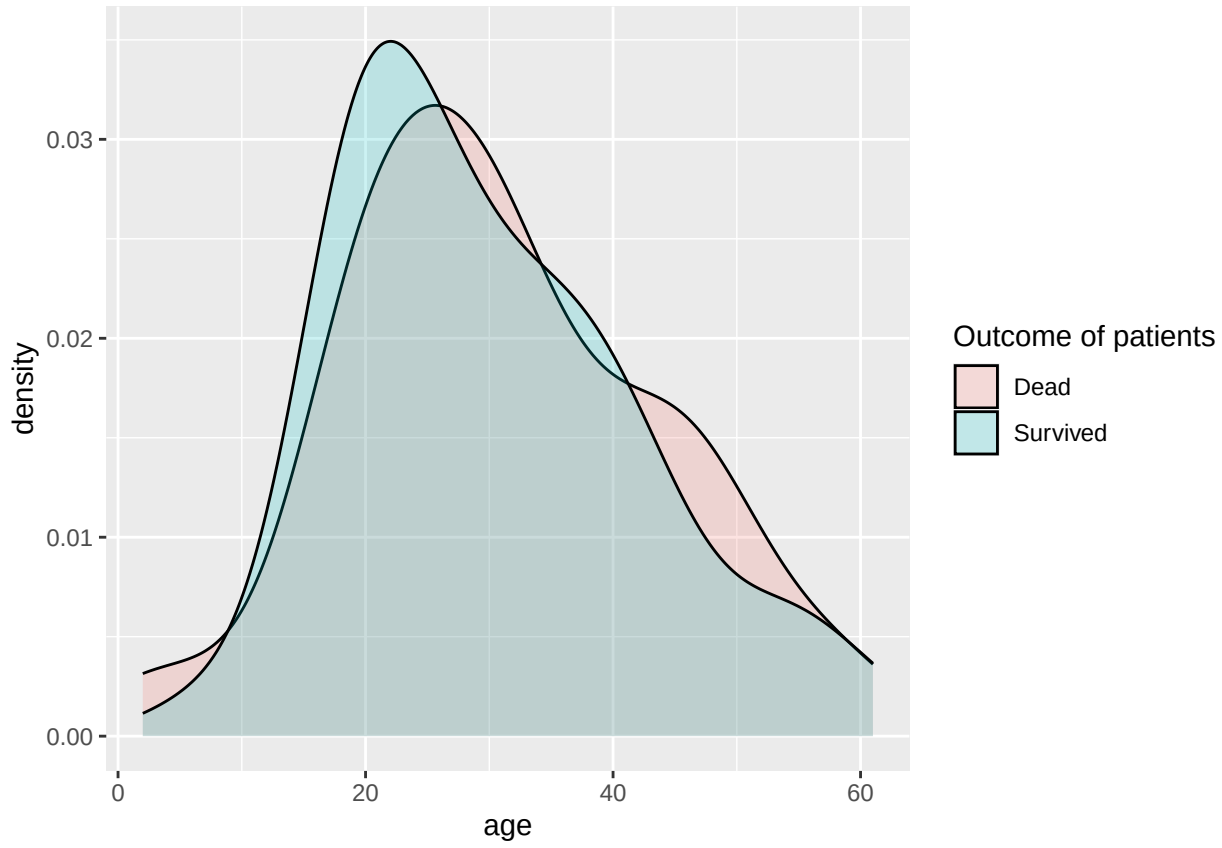


group survived

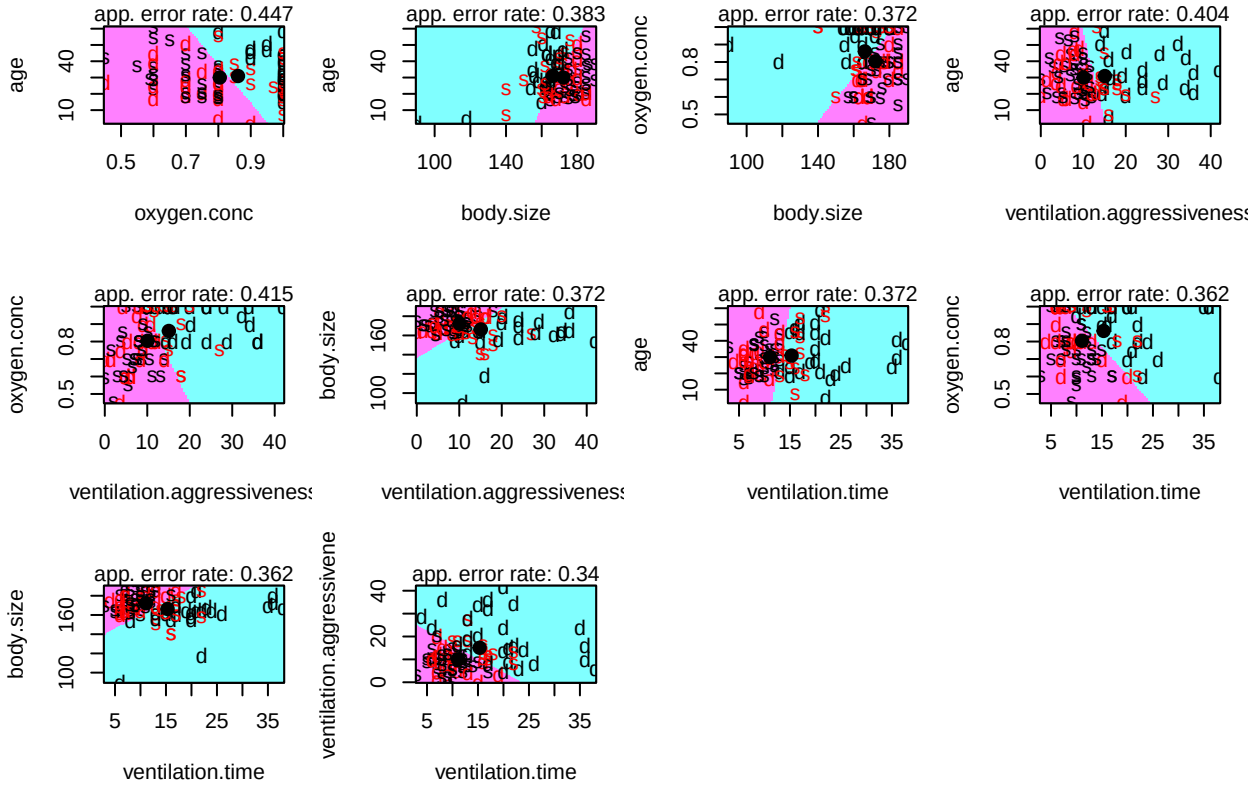
```
lda.data <- cbind(training, predict(linear)$x)
ggplot(lda.data, aes(LD1, age)) +
  geom_point(aes(color = outcome)) +
  labs(title="Outcome of patients with pulmonary problems")
```




```
# plot the classes using density plot
ggplot(data = lda.data)+
  geom_density(aes(age, fill = outcome), alpha = 0.2)+
  scale_fill_discrete(name = "Outcome of patients", labels = c("Dead", "Survived"))
```




```
partimat(factor(outcome)~., data = training,
         method = "lda")
```




```
# prediction accuracy of training set
p1 <- predict(linear, training)$class
tab <- table(Predicted = p1,
             Actual = training$outcome)
tab
```

```
##           Actual
## Predicted  dead survived
##   dead       25        8
##   survived   22       39
```

```
sum(diag(tab))/sum(tab)
```

```
## [1] 0.6808511
```

```
# prediction accuracy of test set
p2 <- predict(linear, testing)$class
tab1 <- table(Predicted = p2,
              Actual = testing$outcome)
tab1
```

```
##           Actual
## Predicted  dead survived
##   dead       11        7
##   survived    5       14
```

```r
sum(diag(tab1))/sum(tab1)
```

```
## [1] 0.6756757
```

```r
# QDA
quadratic <- qda(outcome~., data = training)
quadratic
```

```
## Call:
## qda(outcome ~ ., data = training)
##
## Prior probabilities of groups:
##     dead survived
##      0.5      0.5
##
## Group means:
##               age oxygen.conc body.size ventilation.aggressiveness
## dead     30.93617   0.8606383  166.2340                    15.07598
## survived 30.00000   0.8044681  172.1702                    10.06594
##          ventilation.time
## dead              15.31915
## survived          11.08511
```

```r
predquad <- predict(quadratic, training)
names(predquad)
```

```
## [1] "class"     "posterior"
```

```r
# prediction accuracy of training set
pq1 <- predict(quadratic, training)$class
tab <- table(Predicted = pq1,
               Actual = training$outcome)
tab
```

```
##           Actual
## Predicted  dead survived
##   dead       28        5
##   survived   19       42
```

```r
sum(diag(tab))/sum(tab)
```

```
## [1] 0.7446809
```

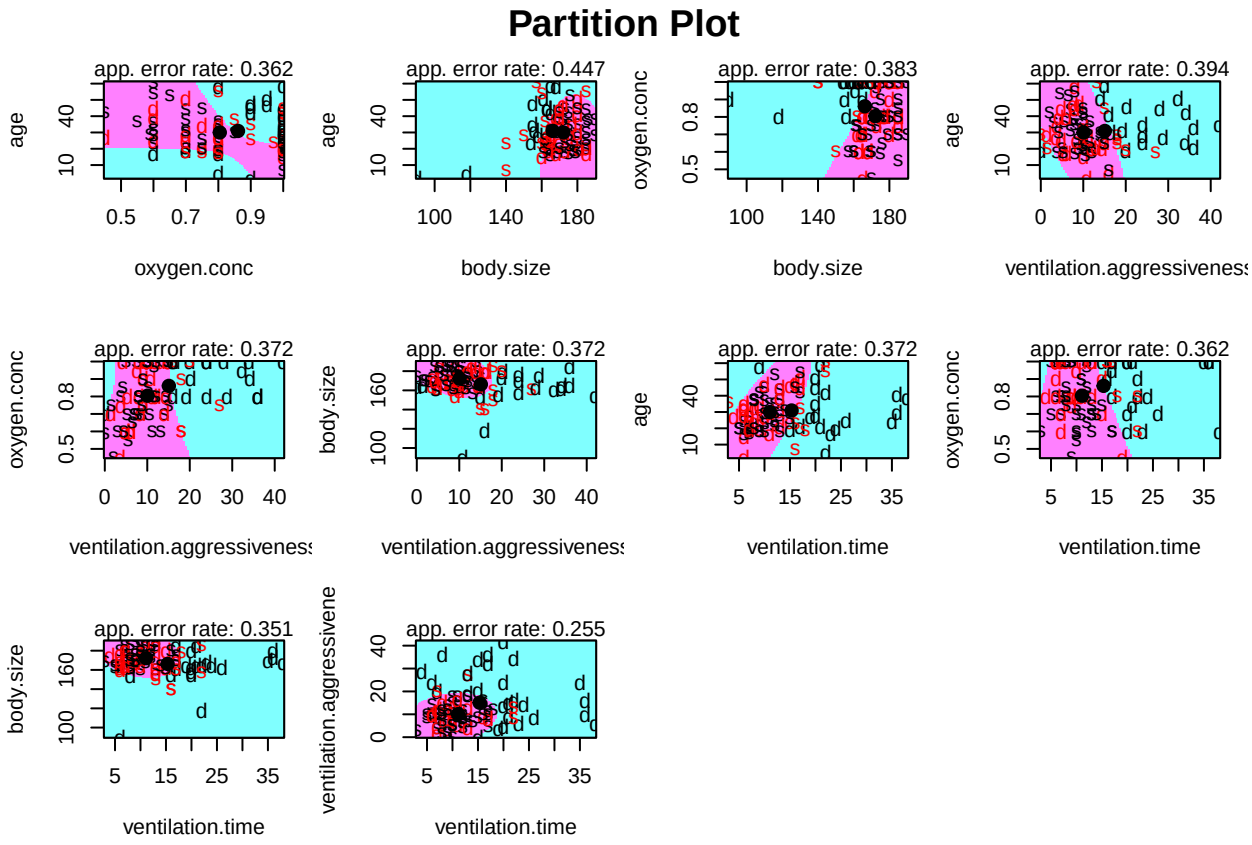
```r
# prediction accuracy of test set
pq2 <- predict(quadratic, testing)$class
tab1 <- table(Predicted = pq2,
                Actual = testing$outcome)
tab1
```

```
##           Actual
## Predicted  dead survived
##   dead       11        7
##   survived    5       14
```

```
sum(diag(tab1))/sum(tab1)
```

```
## [1] 0.6756757
```

```
partimat(factor(outcome)~., data = training,
         method = "qda")
```



From this we can conclude that:

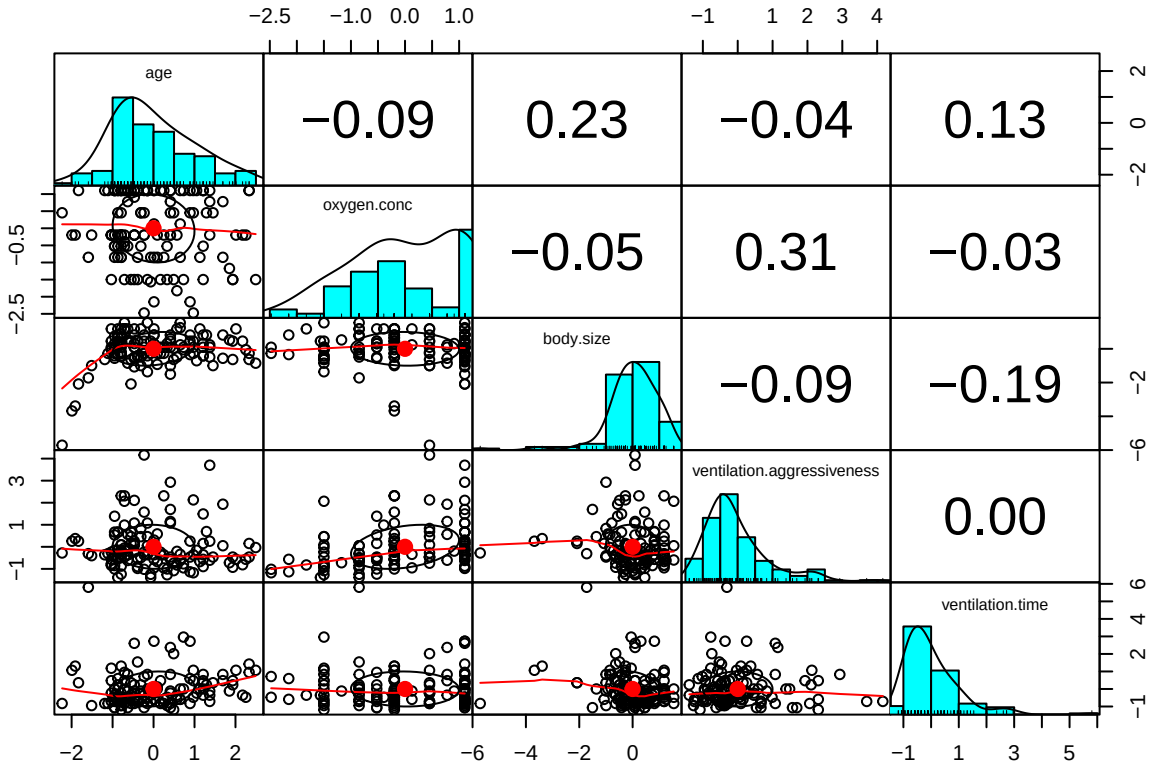
- Prior probabilities of groups are
  - 50% chance to die
  - 50% chance to survive
- LD1 = -0.0085 * age - 2.164 * oxygen.conc + 0.032 * body.size - 0.054 * ventilation.aggressiveness - 0.089 * ventilation.time
- The stacked histogram seems to show us that both groups are overlapping
- However, we can predict with a 68% accuracy on the training set and 67% on the test set, which is a good accuracy despite the overlapping
- Using QDA we get 74% accuracy on the training set and 67% on the test set, which is also an improvement on our learning algorithm.

Now let's try with **standardized** variables:

```
### data with standardization
data2<-read.delim("C:/Users/daria/OneDrive/Desktop/Master - AppDS/Statistics/Datasets-20221007/Discrimi
                  stringsAsFactors=F)

data2_standardized = data2 %>% mutate_at(c('oxygen.conc', 'body.size', 'ventilation.aggressiveness', 'v
                         ~(scale(.) %>% as.vector))

head(data2_standardized)
```

```
##    outcome        age oxygen.conc   body.size ventilation.aggressiveness
## 1     dead -0.2298658  -2.4750733 -0.27082044                  -1.0197847
## 2 survived  0.4108239   0.7783343  0.09193495                   1.1496577
## 3     dead -1.1909002   1.1036751 -0.63357583                  -0.3617548
## 4 survived -0.8705554  -0.5230287  0.45469034                  -0.5416849
## 5 survived -0.7103830  -0.3928924  1.18020113                   0.9851503
## 6 survived -0.4701244  -0.1976880  0.81744573                   0.0906408
##   ventilation.time
## 1        0.7117041
## 2        2.3702369
## 3       -0.8283621
## 4       -0.7098954
## 5       -0.4729622
## 6       -0.8283621
```

```
pairs.panels(data2_standardized[2:6],
             gap = 0,
             bg = c("red", "green")[data2_standardized$outcome],
             pch = 21)
```




```
# 70% for training, 30% for testing
set.seed(123)
ind <- sample(2, nrow(data2_standardized),
              replace = TRUE,
              prob = c(0.7, 0.3))
training <- data2_standardized[ind==1,]
testing <- data2_standardized[ind==2,]

linear <- lda(outcome~., training)
linear
```

```
## Call:
## lda(outcome ~ ., data = training)
##
## Prior probabilities of groups:
##     dead survived
##      0.5      0.5
##
## Group means:
##                  age oxygen.conc  body.size ventilation.aggressiveness
## dead     0.08536719   0.1968742 -0.1812893                   0.1601430
## survived 0.01039287  -0.1686150  0.2493862                  -0.3549754
##          ventilation.time
## dead            0.2756461
## survived       -0.2259467
##
```

```
## Coefficients of linear discriminants:
##                                    LD1
## age                          -0.1063509
## oxygen.conc                  -0.3326207
## body.size                     0.4382951
## ventilation.aggressiveness -0.5291485
## ventilation.time             -0.7540191
```

```
plot(linear)

paste("Prior probabilities of survival: 50%/50% rate of survival")
```

```
## [1] "Prior probabilities of survival: 50%/50% rate of survival"
```

```
# prediction
p <- predict(linear, training)
names(p)
```

```
## [1] "class"     "posterior" "x"
```

```
# Predicted classes
head(p$class, 6)
```

```
## [1] survived survived survived survived survived survived
## Levels: dead survived
```

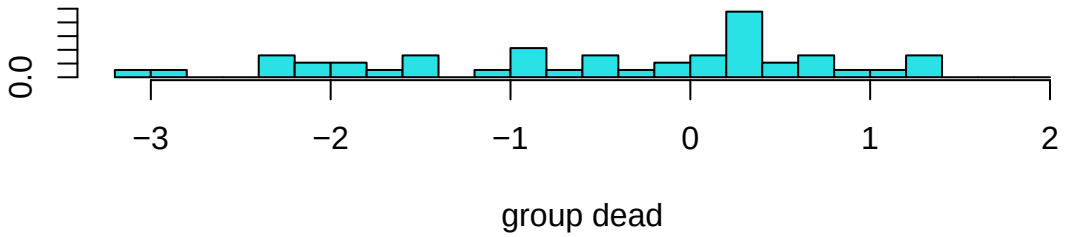
```
# Predicted probabilities
head(p$posterior, 6)
```

```
##         dead  survived
## 1  0.3379166 0.6620834
## 3  0.4373556 0.5626444
## 6  0.2726083 0.7273917
## 7  0.4374854 0.5625146
## 9  0.3881250 0.6118750
## 10 0.4460870 0.5539130
```

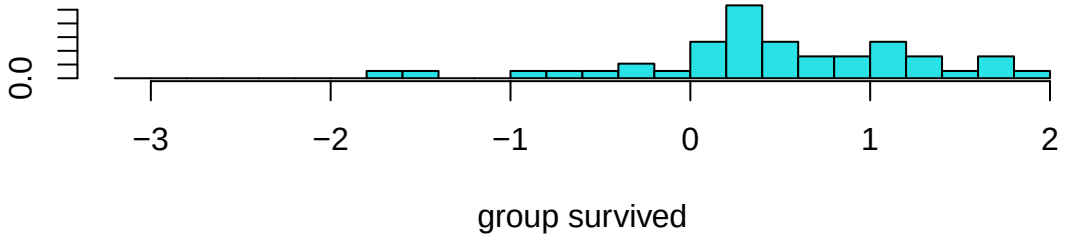
```
# Linear discriminants
head(p$x, 3)
```

```
##         LD1
## 1 0.6940451
## 3 0.2599359
## 6 1.0127322
```

```
# Stacked histogram for LD1
p <- predict(linear, training)
ldahist(data = p$x[,1], g = training$outcome)
```

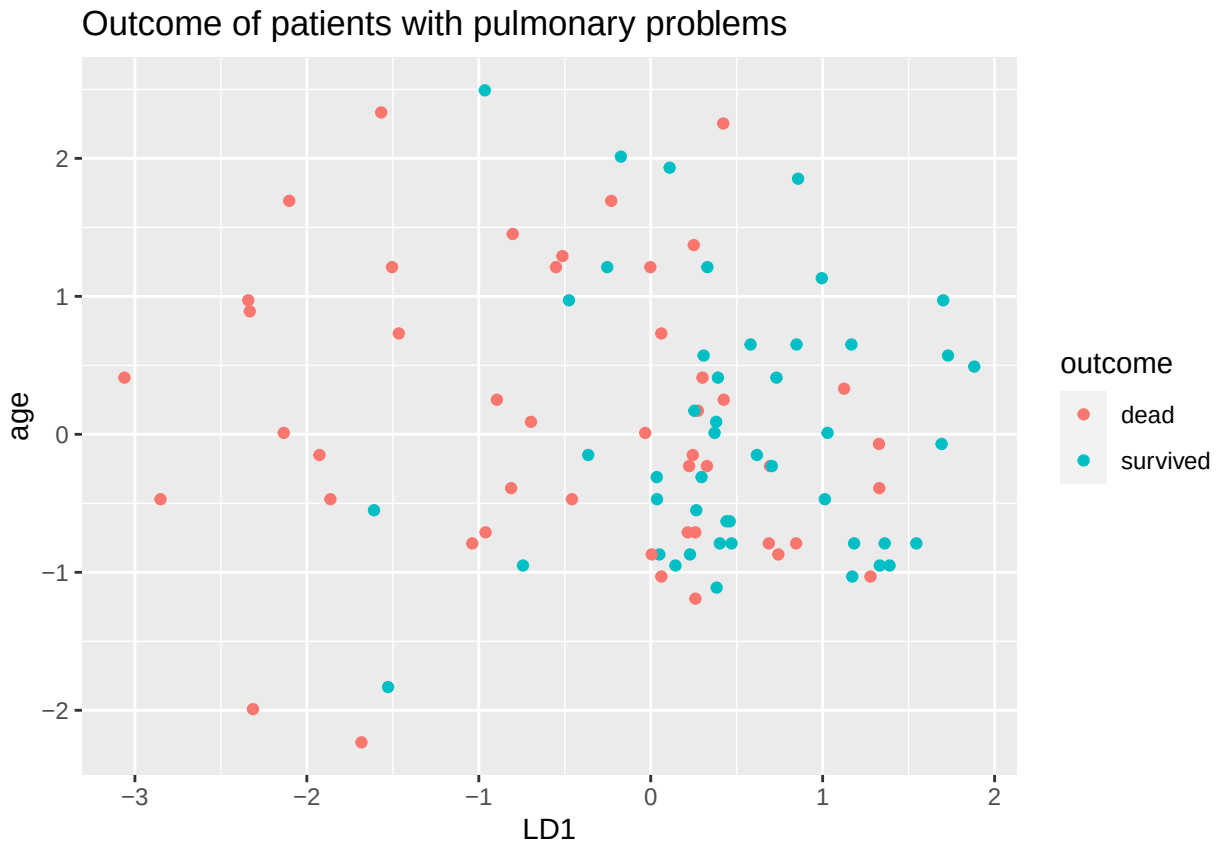


group dead

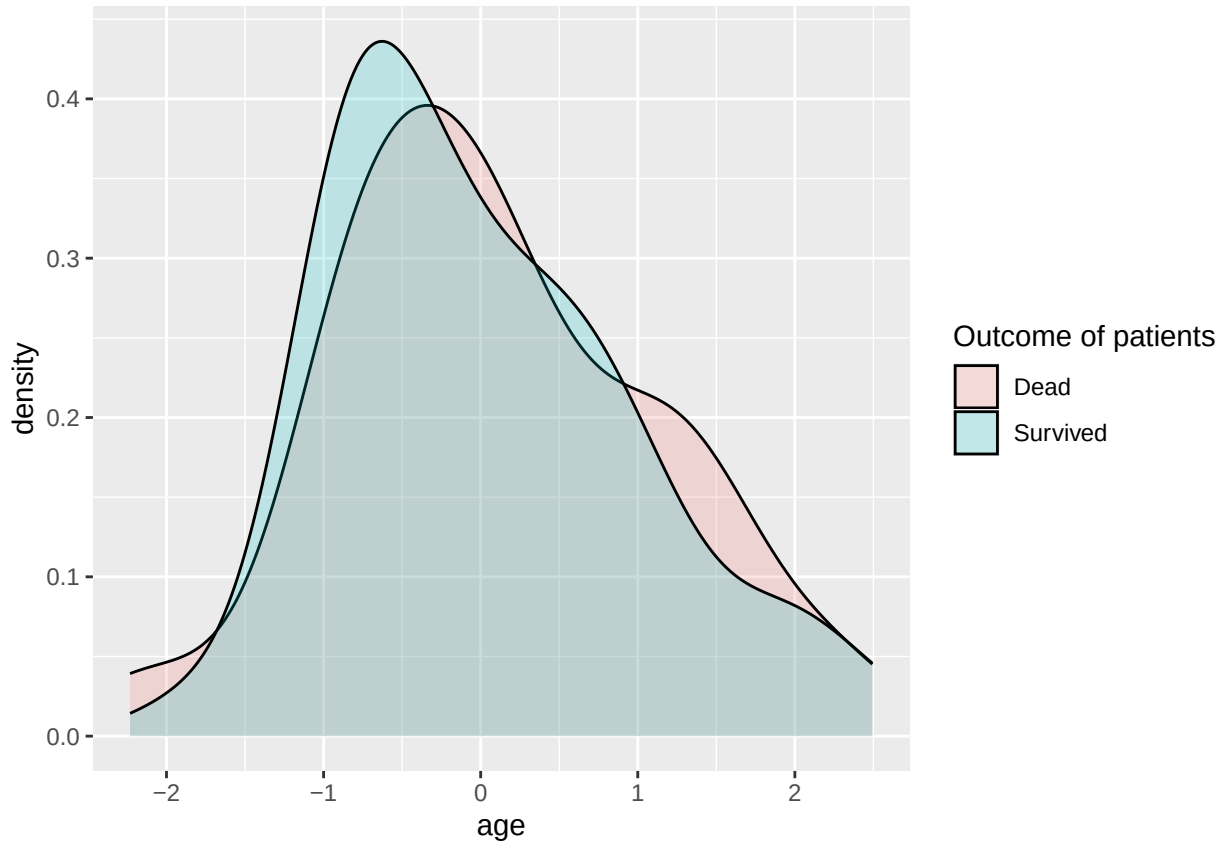


group survived

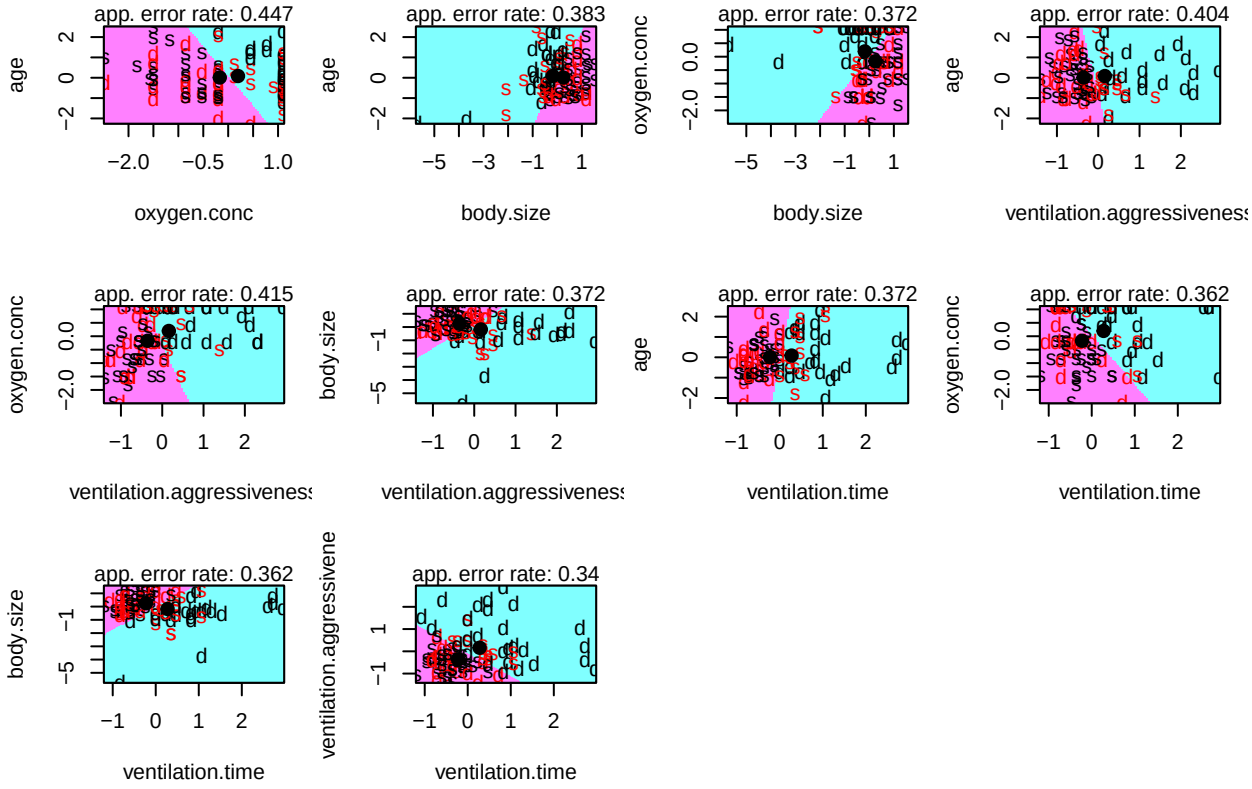
```
lda.data <- cbind(training, predict(linear)$x)
ggplot(lda.data, aes(LD1, age)) +
  geom_point(aes(color = outcome)) +
  labs(title="Outcome of patients with pulmonary problems")
```




```
# plot the classes using density plot
ggplot(data = lda.data)+
  geom_density(aes(age, fill = outcome), alpha = 0.2)+
  scale_fill_discrete(name = "Outcome of patients", labels = c("Dead", "Survived"))
```




```
partimat(factor(outcome)~., data = training,
         method = "lda")
```




```
# prediction accuracy of training set
p1 <- predict(linear, training)$class
tab <- table(Predicted = p1,
             Actual = training$outcome)
tab
```

```
##           Actual
## Predicted  dead survived
##   dead       25        8
##   survived   22       39
```

```
sum(diag(tab))/sum(tab)
```

```
## [1] 0.6808511
```

```
# prediction accuracy of test set
p2 <- predict(linear, testing)$class
tab1 <- table(Predicted = p2,
              Actual = testing$outcome)
tab1
```

```
##           Actual
## Predicted  dead survived
##   dead       11        7
##   survived    5       14
```

```r
sum(diag(tab1))/sum(tab1)
```

```
## [1] 0.6756757
```

```r
# QDA
quadratic <- qda(outcome~., data = training)
quadratic
```

```
## Call:
## qda(outcome ~ ., data = training)
##
## Prior probabilities of groups:
##     dead survived
##      0.5      0.5
##
## Group means:
##                 age oxygen.conc  body.size ventilation.aggressiveness
## dead     0.08536719   0.1968742 -0.1812893                  0.1601430
## survived 0.01039287  -0.1686150  0.2493862                 -0.3549754
##          ventilation.time
## dead            0.2756461
## survived       -0.2259467
```

```r
predquad <- predict(quadratic, training)
names(predquad)
```

```
## [1] "class"     "posterior"
```

```r
# prediction accuracy of training set
pq1 <- predict(quadratic, training)$class
tab <- table(Predicted = pq1,
             Actual = training$outcome)
tab
```

```
##           Actual
## Predicted  dead survived
##   dead       28        5
##   survived   19       42
```

```r
sum(diag(tab))/sum(tab)
```

```
## [1] 0.7446809
```

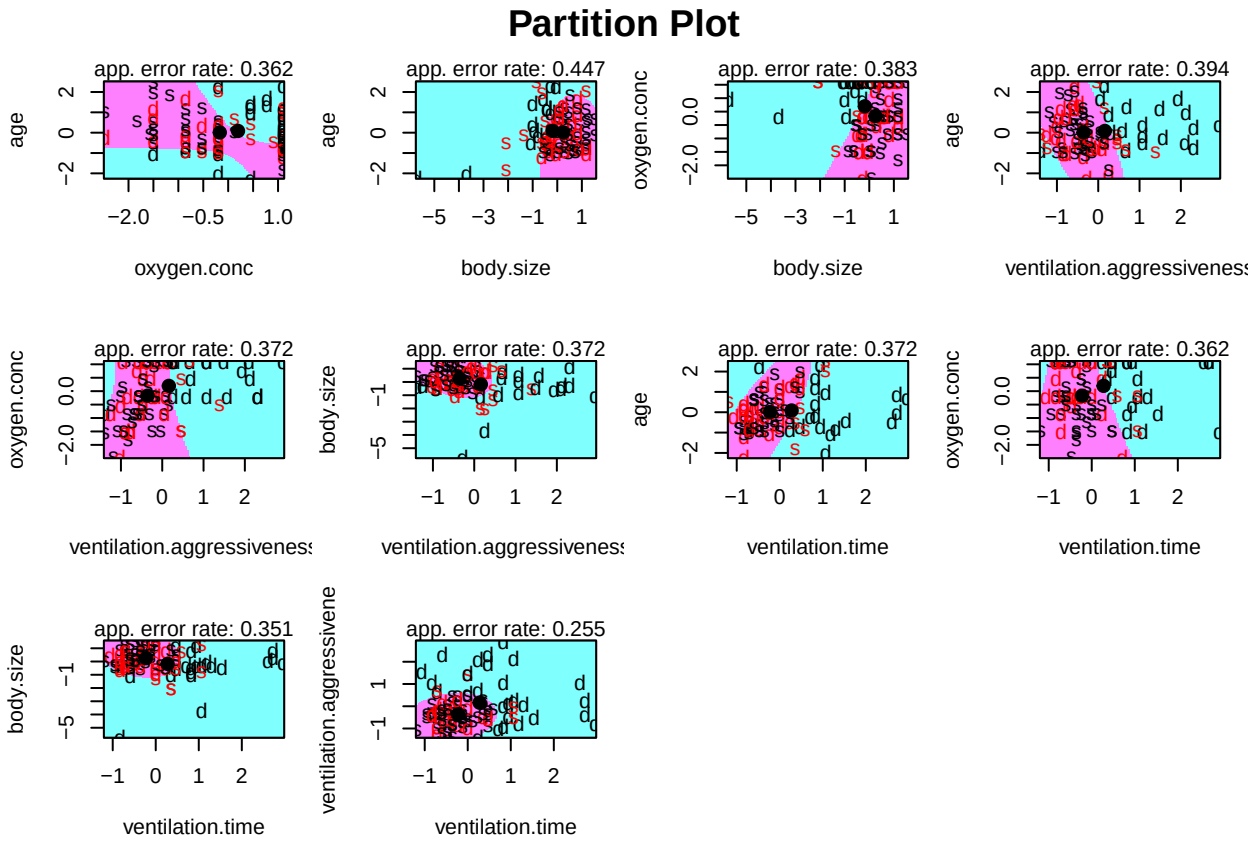
```r
# prediction accuracy of test set
pq2 <- predict(quadratic, testing)$class
tab1 <- table(Predicted = pq2,
              Actual = testing$outcome)
tab1
```

```
##           Actual
## Predicted  dead survived
##   dead       11        7
##   survived    5       14
```

```
sum(diag(tab1))/sum(tab1)
```

```
## [1] 0.6756757
```

```
partimat(factor(outcome)~., data = training,
         method = "qda")
```



We conclude from this standardization that:

- Both not standardized and standardized methods show the same accuracy on predictibility
- Standardization is better for understanding the importance of our data

Exercise 86

- In bird species where males and females are coloured the same, it is usually difficult to determine sex on the basis of external characteristics such as size or behaviour.
- Therefore, either an endoscopic examination of the internal reproductive organs or a blood sample must be taken.

- Both procedures put the animals under a lot of stress, so they usually take place under general anaesthesia, which can also affect the birds' health.
- In addition, analysing the blood for hormone status or certain genetic traits is costly and requires specialist staff.
- The aim of this study is to determine the sex of the birds based on the characteristics of wing length, beak length, head length, foot length and weight.
- The aim of this study is to establish a discriminant function for sex determination using data from 245 birds, most of which have known sex.
- The following research questions need to be answered:
  - What is the discriminant function?
  - How many birds can be correctly assigned to their sex with this?
  - How many of the birds whose sex is unknown can still be assigned to a sex?

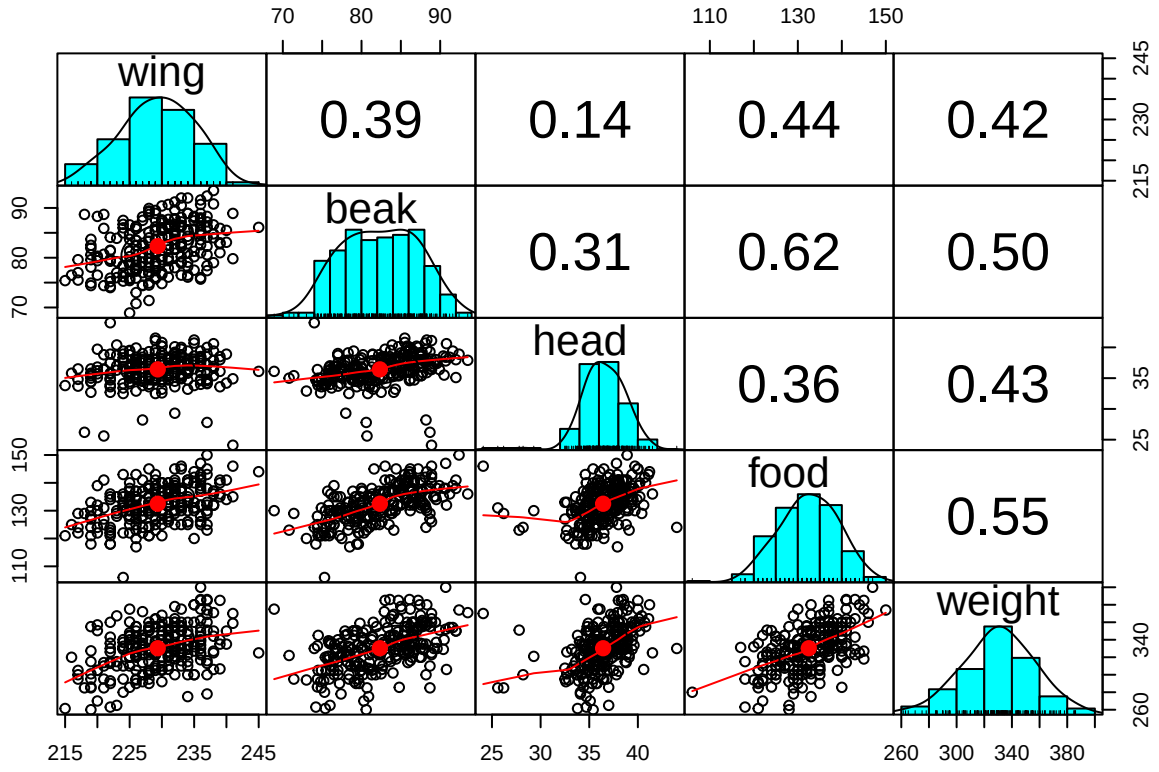
```
#### Exercise 86 ####
data3<-read.delim("C:/Users/daria/OneDrive/Desktop/Master - AppDS/Statistics/Datasets-20221007/Discrimi
                  stringsAsFactors=F)

# Look at the data
head(data3)
```

```
##   ï..sex wing beak head food weight
## 1   male  226 85.8 39.9  134  355.0
## 2   male  226 86.2 34.8  137  355.0
## 3   male  238 84.8 35.4  137  327.0
## 4   male  231 79.4 36.7  126  310.0
## 5   male  230 80.3 35.8  130  340.0
## 6   male  220 79.4 35.7  124  350.4
```

```
# Get type of the data
str(data3)
```

```
## 'data.frame':    245 obs. of  6 variables:
##  $ ï..sex: chr  "male" "male" "male" "male" ...
##  $ wing  : int  226 226 238 231 230 220 225 230 225 230 ...
##  $ beak  : num  85.8 86.2 84.8 79.4 80.3 79.4 79.1 77 78.6 81.5 ...
##  $ head  : num  39.9 34.8 35.4 36.7 35.8 35.7 35.7 37.5 36 33.2 ...
##  $ food  : int  134 137 137 126 130 124 127 130 128 129 ...
##  $ weight: num  355 355 327 310 340 ...
```

```
pairs.panels(data3[2:6],
             gap = 0,
             bg = c("red", "green")[data3$ï..sex],
             pch = 21)
```




```
# 70% for training, 30% for testing
set.seed(123)
ind <- sample(2, nrow(data3),
              replace = TRUE,
              prob = c(0.7, 0.3))
training <- data3[ind==1,]
testing <- data3[ind==2,]

linear <- lda(ï..sex~., training)
linear
```

```
## Call:
## lda(ï..sex ~ ., data = training)
##
## Prior probabilities of groups:
##    female      male   unknown
## 0.3222222 0.1555556 0.5222222
##
## Group means:
##              wing     beak     head     food   weight
## female  231.2241 86.92586 38.20172 137.1897 347.2483
## male    227.2143 77.71786 35.20000 128.8571 319.0357
## unknown 228.7660 80.68085 35.76277 130.6596 323.3000
##
## Coefficients of linear discriminants:
##                 LD1         LD2
```

```
## wing    0.026379351  0.09634064
## beak   -0.259704762  0.18016711
## head   -0.211730927 -0.10456727
## food    0.013755302 -0.08413990
## weight -0.004460849 -0.02953404
##
## Proportion of trace:
##    LD1    LD2
## 0.9875 0.0125
```

```
paste("Prior probabilities of sex: 32% female, 15% male, 52% unknown")
```

```
## [1] "Prior probabilities of sex: 32% female, 15% male, 52% unknown"
```

```
# prediction
p <- predict(linear, training)
names(p)
```

```
## [1] "class"     "posterior" "x"
```

```
# Predicted classes
head(p$class, 6)
```

```
## [1] female  unknown unknown unknown unknown unknown
## Levels: female male unknown
```

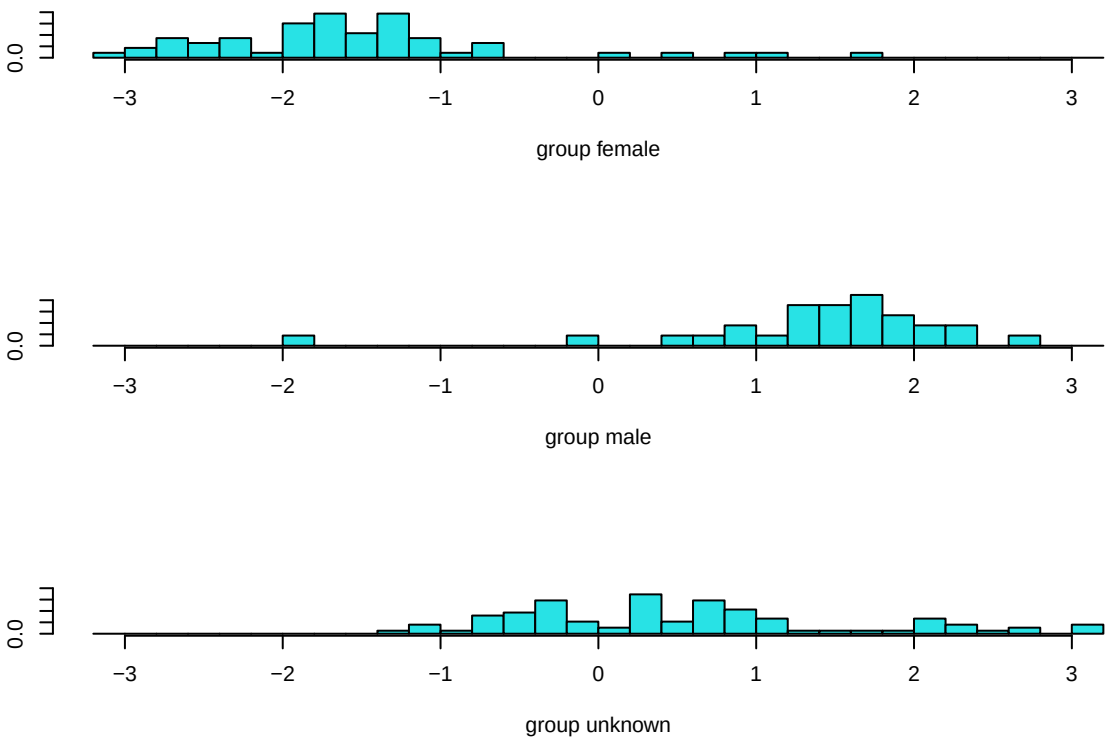
```
# Predicted probabilities
head(p$posterior, 6)
```

```
##        female        male    unknown
## 1  0.91647194 0.003030302 0.08049776
## 3  0.18405686 0.058914522 0.75702862
## 6  0.07381319 0.205611452 0.72057535
## 7  0.03985952 0.232927343 0.72721314
## 9  0.02233897 0.221667057 0.75599397
## 10 0.01988418 0.153214117 0.82690170
```

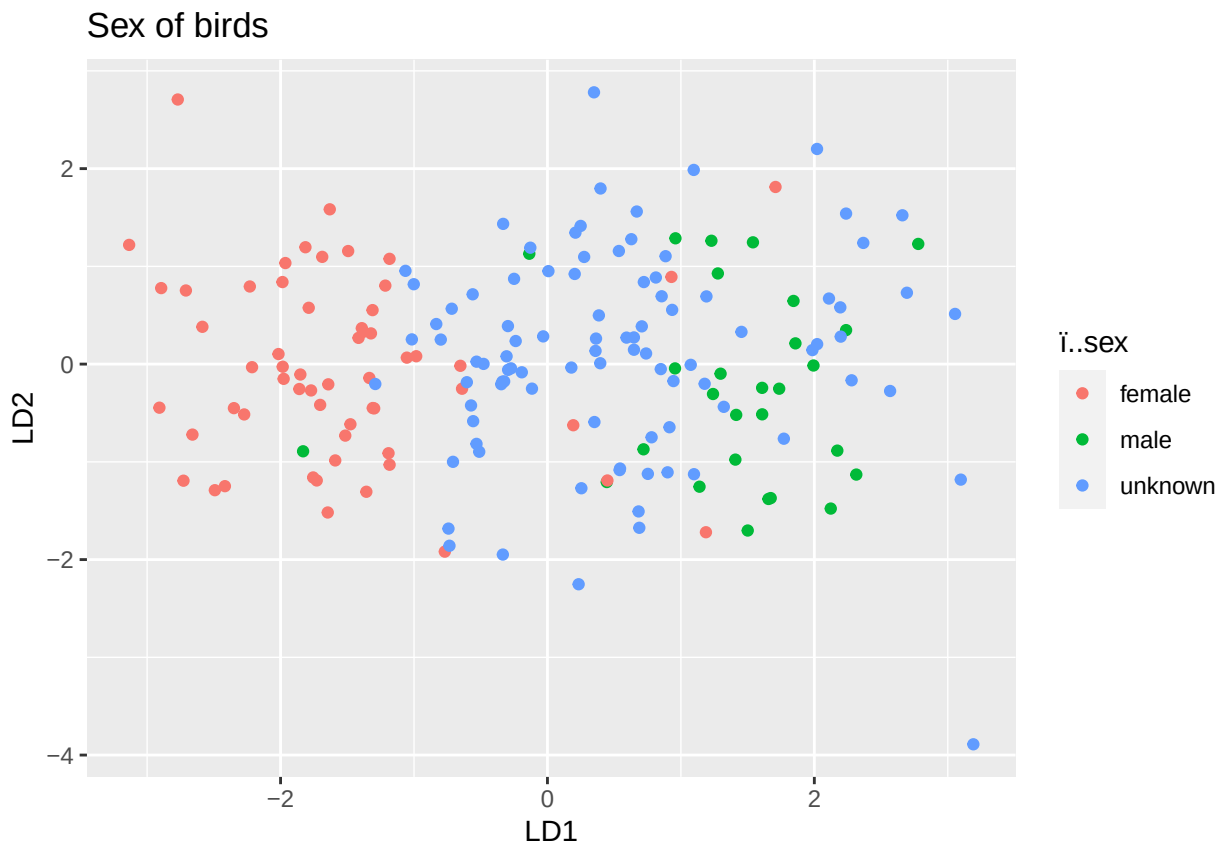
```
# Linear discriminants
head(p$x, 3)
```

```
##          LD1        LD2
## 1 -1.8312624 -0.8918567
## 3 -0.1360466  1.1291501
## 6  0.4448088 -1.2065319
```

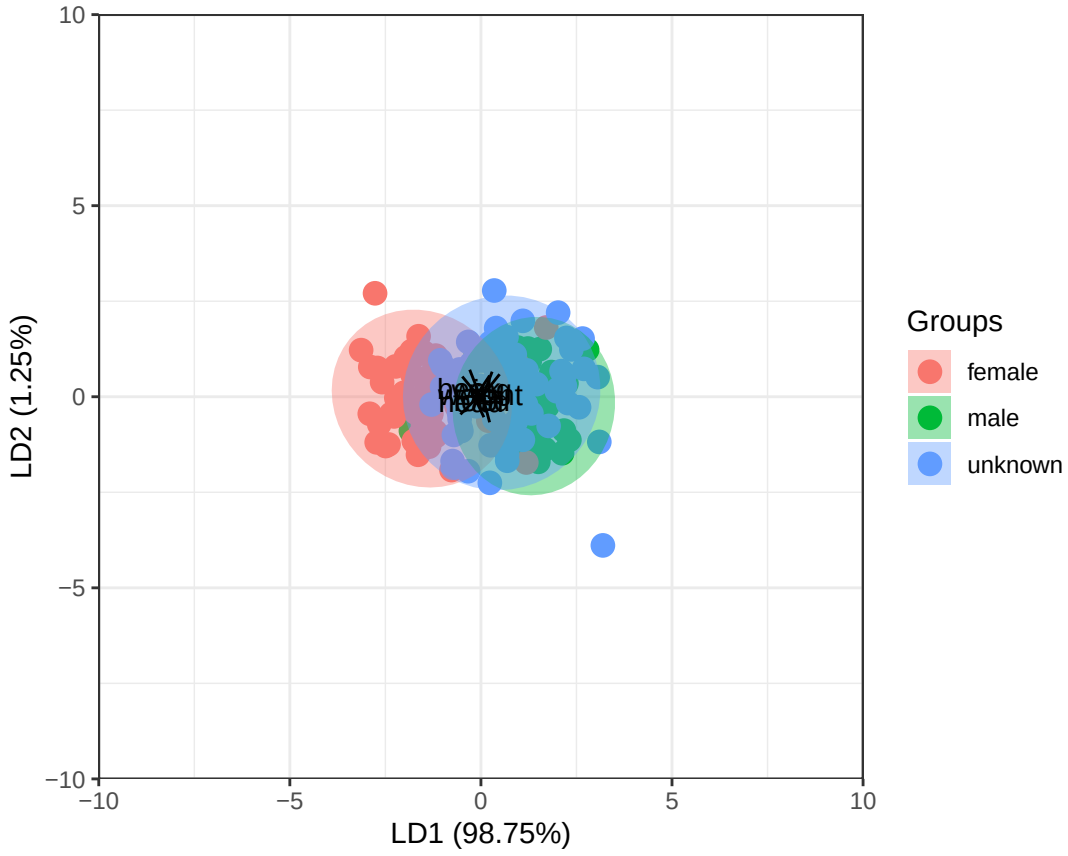
```
# Stacked histogram for LD1
p <- predict(linear, training)
ldahist(data = p$x[,1], g = training$ï..sex)
```



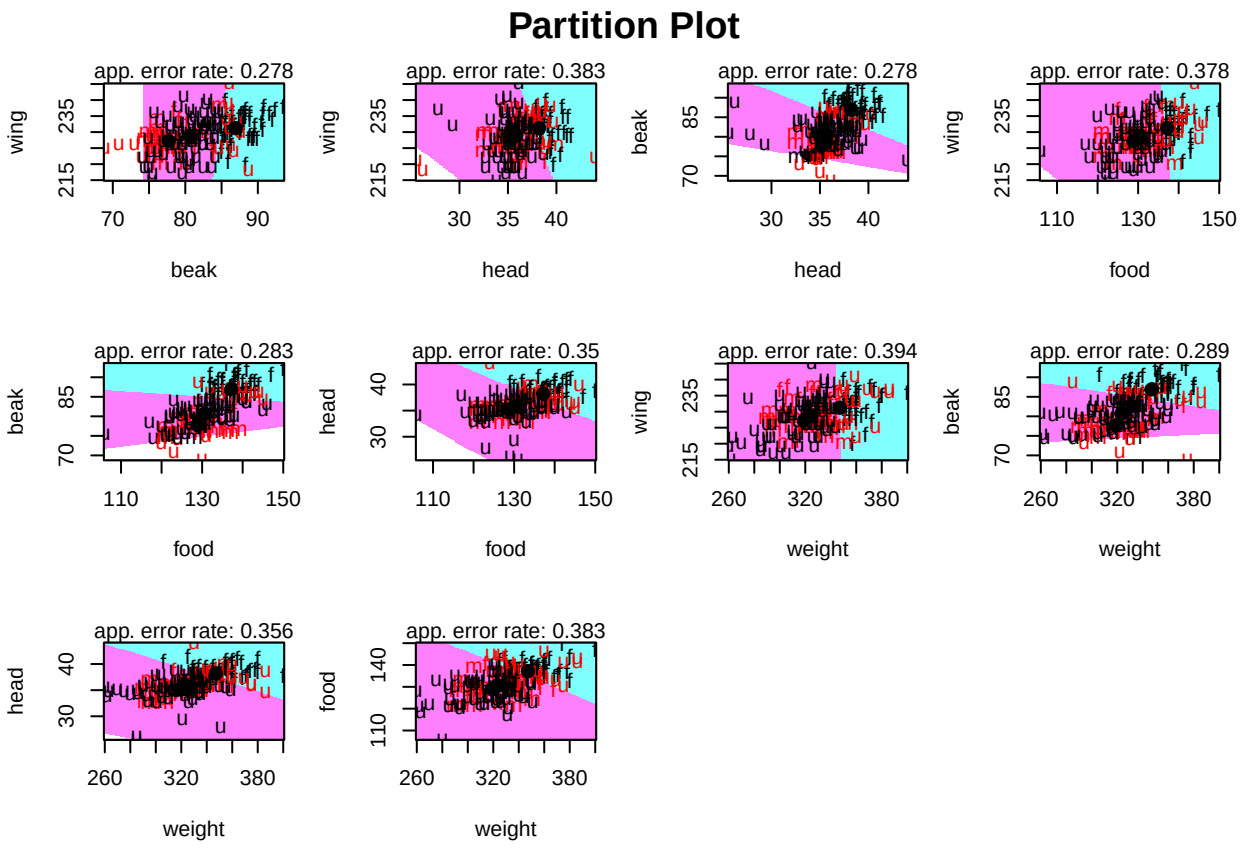

```
lda.data <- cbind(training, predict(linear)$x)
ggplot(lda.data, aes(LD1, LD2)) +
  geom_point(aes(color = ï..sex)) +
  labs(title="Sex of birds")
```




```
# Biplot
ggord(linear, training$ï..sex,
      ylim = c(-10, 10), xlim=c(-10,10))
```




```
partimat(factor(ï..sex)~., data = training,
         method = "lda")
```




```
# prediction accuracy of training set
p1 <- predict(linear, training)$class
tab <- table(Predicted = p1,
             Actual = training$ï..sex)
tab
```

```
##           Actual
## Predicted female male unknown
##   female      51    1       9
##   male         0    3       4
##   unknown      7   24      81
```

```
sum(diag(tab))/sum(tab)
```

```
## [1] 0.75
```

```
# prediction accuracy of test set
p2 <- predict(linear, testing)$class
tab1 <- table(Predicted = p2,
              Actual = testing$ï..sex)
tab1
```

```
##           Actual
## Predicted female male unknown
```

```
##   female      20    1       5
##   male         0    0       2
##   unknown      2   13      22
```

```
sum(diag(tab1))/sum(tab1)
```

```
## [1] 0.6461538
```

```
# QDA
quadratic <- qda(ï..sex~., data = training)
quadratic
```

```
## Call:
## qda(ï..sex ~ ., data = training)
##
## Prior probabilities of groups:
##    female      male   unknown
## 0.3222222 0.1555556 0.5222222
##
## Group means:
##             wing     beak     head     food   weight
## female  231.2241 86.92586 38.20172 137.1897 347.2483
## male    227.2143 77.71786 35.20000 128.8571 319.0357
## unknown 228.7660 80.68085 35.76277 130.6596 323.3000
```

```
predquad <- predict(quadratic, training)
names(predquad)
```

```
## [1] "class"     "posterior"
```

```
# prediction accuracy of training set
pq1 <- predict(quadratic, training)$class
tab <- table(Predicted = pq1,
             Actual = training$ï..sex)
tab
```

```
##          Actual
## Predicted female male unknown
##   female      52    1      10
##   male         2   18       7
##   unknown      4    9      77
```

```
sum(diag(tab))/sum(tab)
```

```
## [1] 0.8166667
```

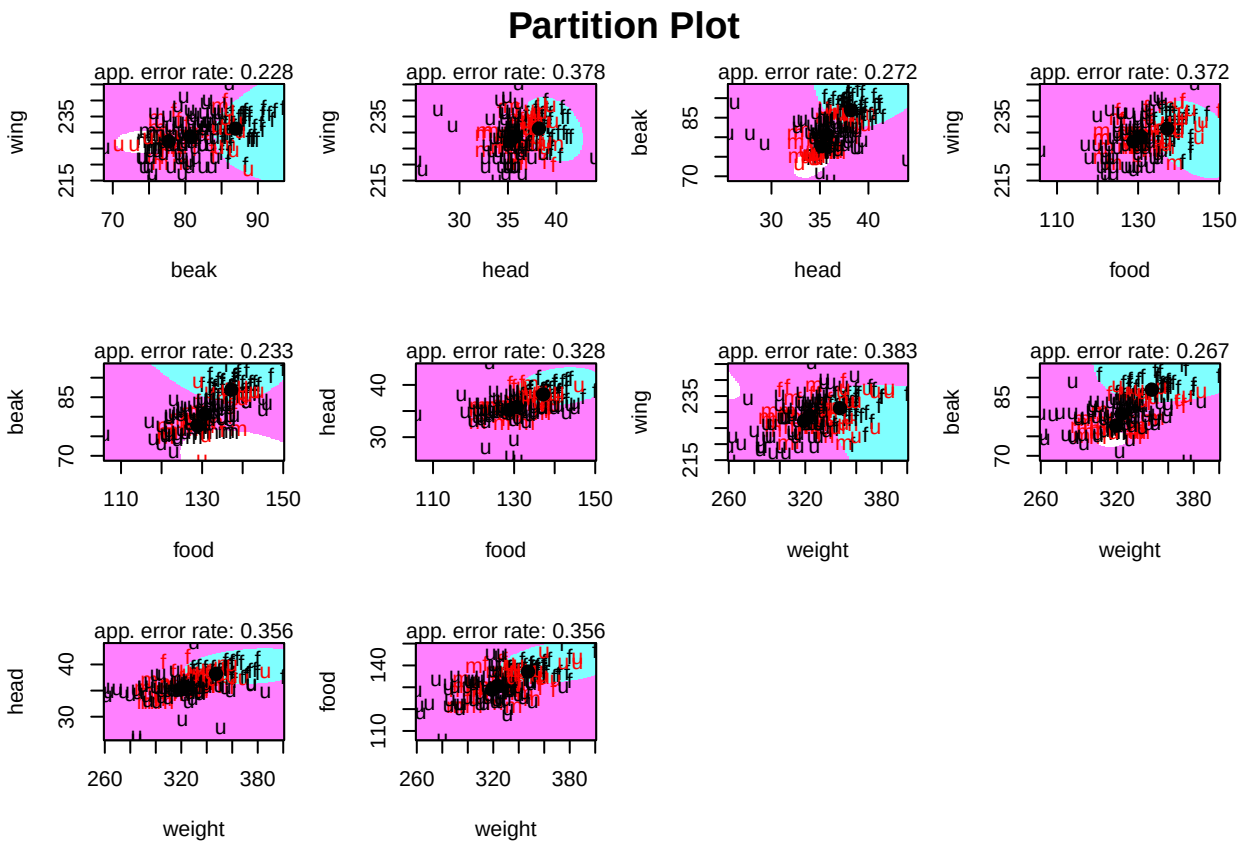
```
# prediction accuracy of test set
pq2 <- predict(quadratic, testing)$class
tab1 <- table(Predicted = pq2,
              Actual = testing$ï..sex)
tab1
```

```
##           Actual
## Predicted female male unknown
##   female      17    0       5
##   male         1    4       4
##   unknown      4   10      20
```

```
sum(diag(tab1))/sum(tab1)
```

```
## [1] 0.6307692
```

```
partimat(factor(ï..sex)~., data = training,
         method = "qda")
```



Thus, we can conclude that:

- Prior probabilities of sex groups are:
  - 32% female
  - 16% male
  - 52% unknown
- The discriminant functions are:
  - LD1 = 0.026 * wing - 0.260 * beak - 0.212 * head + 0.014 * food - 0.004 * weight
  - LD2 = 0.096 * wing + 0.180 * beak - 0.105 * head - 0.084 * food - 0.030 * weight
- There are quite no overlaps of the stacked histograms

- The scatter plot shows a better performance for QDA method
- There is 75% accuracy on the trained test and 64% accuracy on the test set, thus, a 64% chance to assign the sex correctly to a bird using LDA method
- There is 82% accuracy on the trained test and 63% accuracy on the test set, thus, a 63% chance to assign the sex correctly to a bird using QDA method
- There is a total of 123 birds whose sex is still unknown and still can be assigned. [Total sum of unknown column for both training and test set]