Project 3: Discriminant Analysis

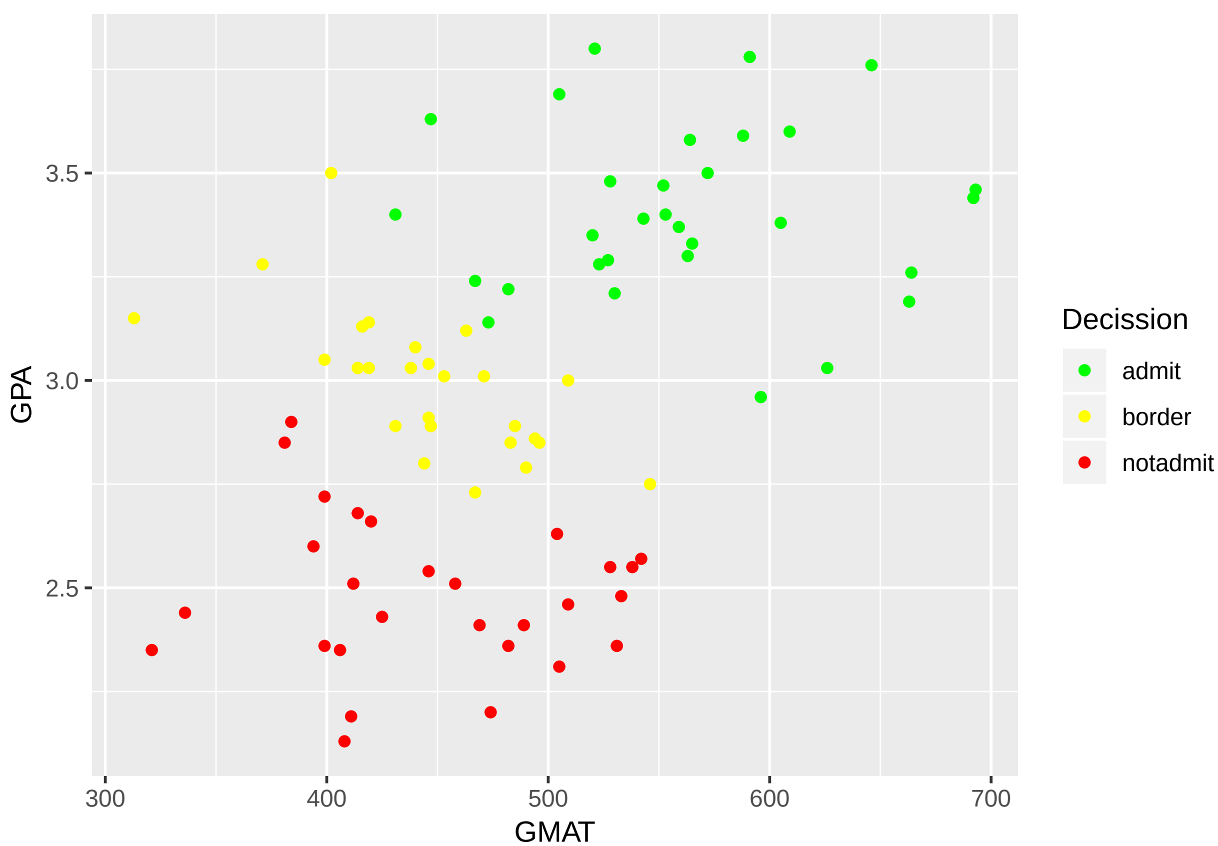
DA 410

Marjorie Blanco

Problem 1

1. Use admission.csv as a training dataset.

GPA	GMAT	Decission
2.96	596	admit
3.14	473	admit
3.22	482	admit
3.29	527	admit
3.69	505	admit




```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  train.admission[, 1:2]
## Chi-Sq (approx.) = 16.074, df = 6, p-value = 0.01336
```

The results indicate that the three groups have similar variance-covariance matrices.

2. Train model using LDA by setting admit/not-admit/border with the same probabilities.

The purpose of linear discriminant analysis (LDA) is to find the linear combinations of the original variables (GPA and GMAT) that gives the best possible separation between the groups (admission recomendation) in the data set.

```r
# Fit the model
lda.model1 <- lda(Decission~., data = train.admission)
lda.model1
```

```
## Call:
## lda(Decission ~ ., data = train.admission)
##
## Prior probabilities of groups:
##     admit    border  notadmit
## 0.3647059 0.3058824 0.3294118
##
## Group means:
##               GPA     GMAT
## admit    3.403871 561.2258
## border   2.992692 446.2308
## notadmit 2.482500 447.0714
##
## Coefficients of linear discriminants:
##              LD1         LD2
## GPA  5.008766354  1.87668220
## GMAT 0.008568593 -0.01445106
##
## Proportion of trace:
##    LD1    LD2
## 0.9673 0.0327
```

The first discriminant function is a linear combination of the variables: $5.008766354xGPA + 0.008568593xGMAT$

The second discriminant function is a linear combination of the variables: $1.87668220xGPA - 0.01445106xGMAT$

The LDA probability of admitting is 36% while probability of not admitting is 33% and probability of border is 31%.

3. Calculate the misclassfication rate

The training model correctly classified 92.9% of observations, which is an increase from problem 1 accuracy.

The training model misclassification rate for LDA is 7.1%.

Confusion Matrix

```
## Confusion Matrix and Statistics
##
```

```
##           Reference
## Prediction admit border notadmit
##   admit       28      3        0
##   border       1     24        1
##   notadmit     0      2       26
##
## Overall Statistics
##
##                Accuracy : 0.9176
##                  95% CI : (0.8377, 0.9662)
##     No Information Rate : 0.3412
##     P-Value [Acc > NIR] : < 0.00000000000000022
##
##                   Kappa : 0.8765
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: admit Class: border Class: notadmit
## Sensitivity                0.9655        0.8276          0.9630
## Specificity                0.9464        0.9643          0.9655
## Pos Pred Value             0.9032        0.9231          0.9286
## Neg Pred Value             0.9815        0.9153          0.9825
## Prevalence                 0.3412        0.3412          0.3176
## Detection Rate             0.3294        0.2824          0.3059
## Detection Prevalence       0.3647        0.3059          0.3294
## Balanced Accuracy          0.9560        0.8959          0.9642
```

Here, we can see that 28 out of 31 admit decission are expected to be correctly classified, 24 out of 26 border decission are expected to be correctly classified, and 26 out of 28 notadmit decission are expected to be correctly classified.

Chi-squared

Observed:

```
##           Reference
## Prediction admit border notadmit
##   admit       28      3        0
##   border       1     24        1
##   notadmit     0      2       26
```

Expected:

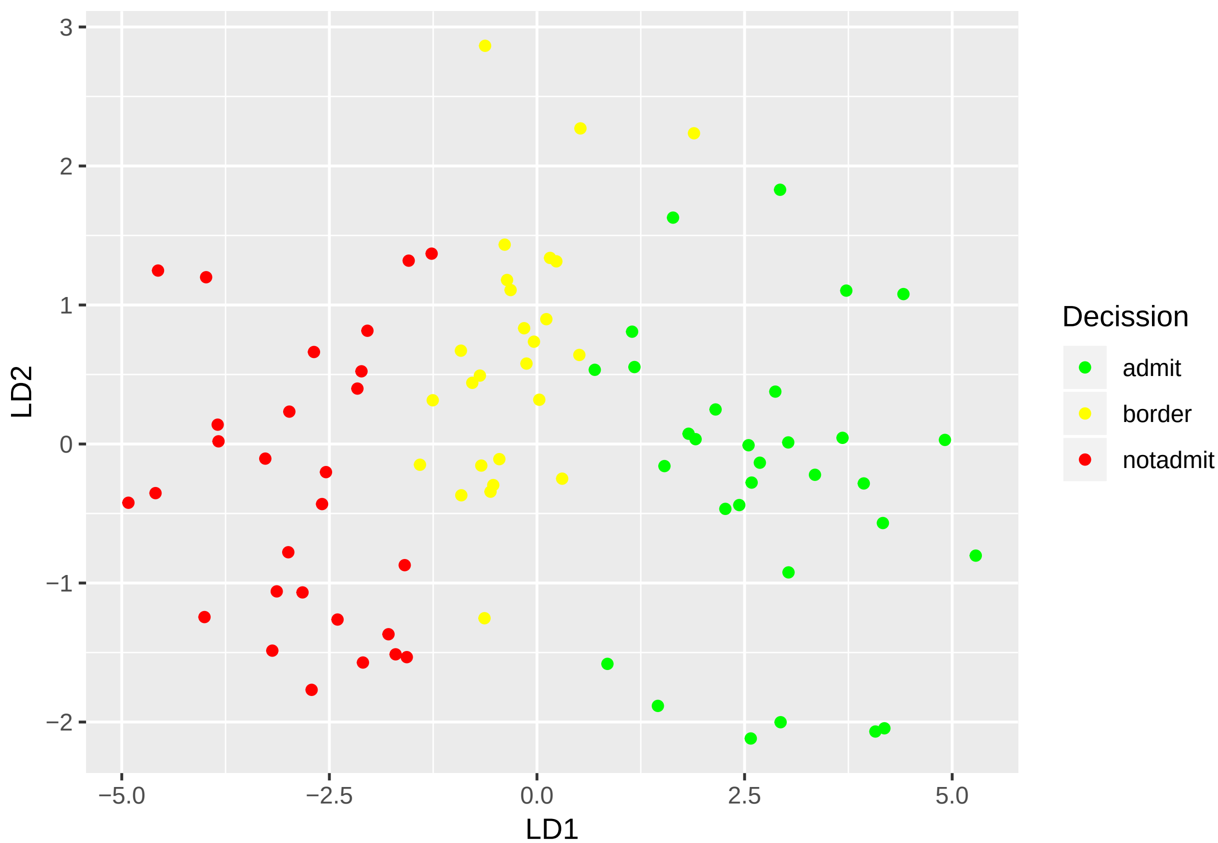
```
##           Reference
## Prediction     admit    border notadmit
##   admit    10.576471 10.576471 9.847059
##   border    8.870588  8.870588 8.258824
##   notadmit  9.552941  9.552941 8.894118
```

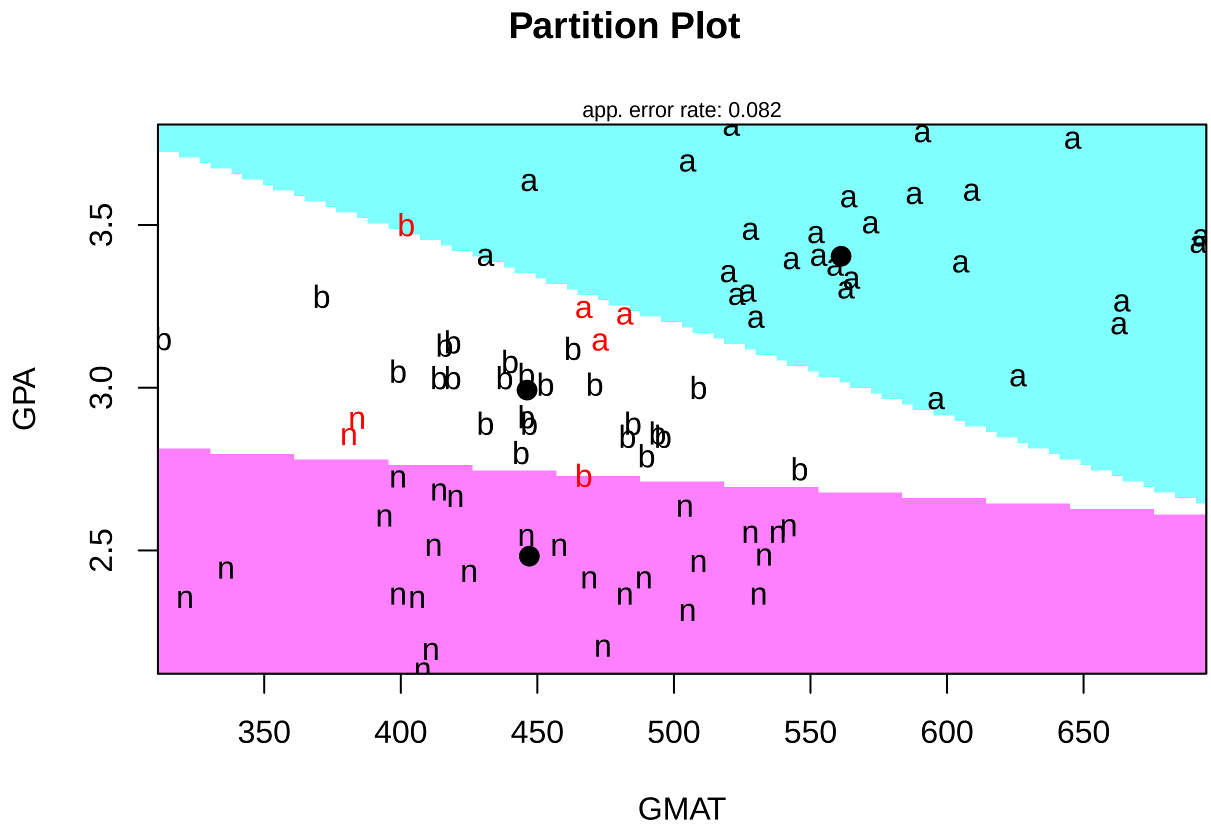
Residuals:

```
##           Reference
## Prediction     admit    border  notadmit
##   admit      5.357544 -2.329682 -3.138002
##   border    -2.642597  5.079791 -2.525848
##   notadmit -3.090783 -2.443698  5.735801
```

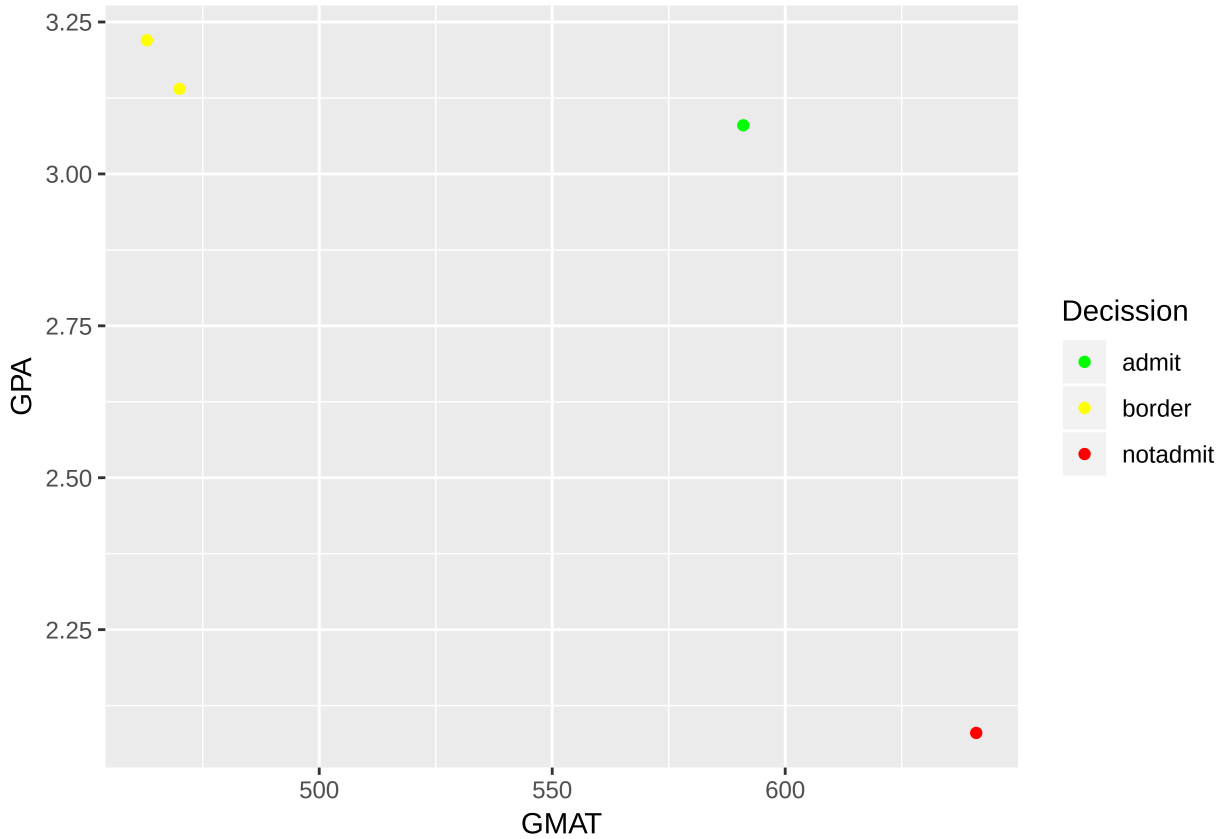
Standardized residuals:

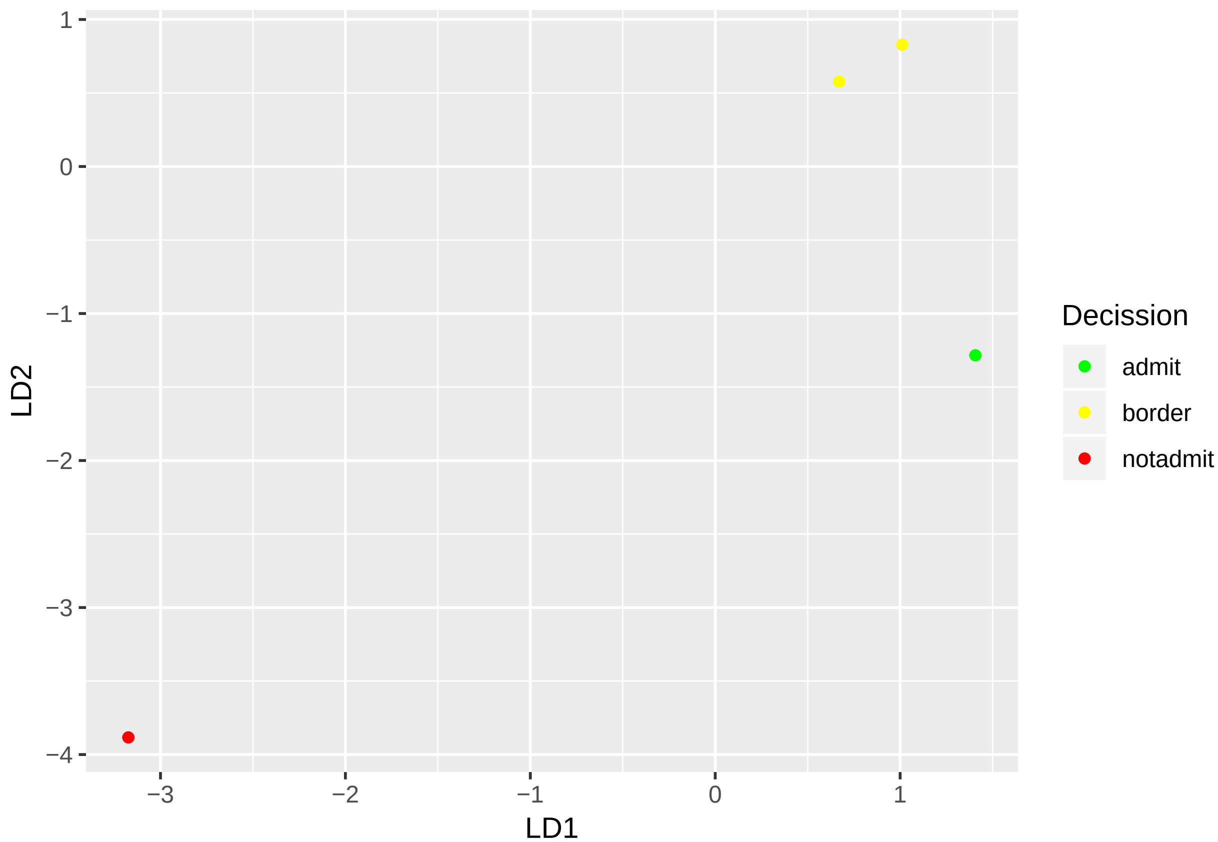
```
##           Reference
## Prediction     admit    border  notadmit
##   admit      8.281210 -3.601012 -4.766080
##   border    -3.907778  7.511813 -3.670169
##   notadmit -4.650033 -3.676504  8.479330
```





4. Predict students with GPA and GMAT score as below.


3.25
3.00
2.75
2.50
2.25
GPA
500
550
600
GMAT
Decission
admit
border
notadmit



GPA	GMAT	LD1	LD2	Decission
3.14	470	0.6704435	0.5770049	border
3.08	591	1.4067173	-1.2841743	admit
2.08	641	-3.1736194	-3.8834095	notadmit
3.22	463	1.0111647	0.8282969	border

Problem 2

1. Use admission.csv as a training dataset.
2. Train model using LDA by setting probability of admit is 50% while probability of not admit is 25% and probability of border is 25%.

```
## Call:
## lda(Decission ~ ., data = train.admission, prior = c(0.5, 0.25,
##     0.25))
##
## Prior probabilities of groups:
##    admit   border notadmit
##     0.50     0.25     0.25
##
## Group means:
##                GPA     GMAT
## admit     3.403871 561.2258
## border    2.992692 446.2308
```

```
## notadmit 2.482500 447.0714
##
## Coefficients of linear discriminants:
##              LD1        LD2
## GPA  4.961868967  1.9973815
## GMAT 0.008915905 -0.0142394
##
## Proportion of trace:
##    LD1    LD2
## 0.9724 0.0276
```

The first discriminant function is a linear combination of the variables: $4.961868967xGPA + 0.008915905xGMAT$

The second discriminant function is a linear combination of the variables: $1.9973815xGPA - 0.0142394xGMAT$

3. Calculate the misclassfication rate

The training model correctly classified 92.9% of observations, which is an increase from problem 1 accuracy.

The training model misclassification rate for LDA is 7.1%.

Confusion Matrix

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction admit border notadmit
##   admit       29      2        0
##   border       1     25        0
##   notadmit     0      2       26
##
## Overall Statistics
##
##                Accuracy : 0.9412
##                  95% CI : (0.868, 0.9806)
##     No Information Rate : 0.3529
##     P-Value [Acc > NIR] : < 0.00000000000000022
##
##                   Kappa : 0.9117
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: admit Class: border Class: notadmit
## Sensitivity                0.9667        0.8621          1.0000
## Specificity                0.9636        0.9821          0.9661
## Pos Pred Value             0.9355        0.9615          0.9286
## Neg Pred Value             0.9815        0.9322          1.0000
## Prevalence                 0.3529        0.3412          0.3059
## Detection Rate             0.3412        0.2941          0.3059
## Detection Prevalence       0.3647        0.3059          0.3294
## Balanced Accuracy          0.9652        0.9221          0.9831
```

Here, we can see that 29 out of 31 admit decission are expected to be correctly classified, 25 out of 26 border decission are expected to be correctly classified, and 26 out of 28 notadmit decission are expected to be correctly classified.

Chi-squared

Observed:

```
##           Reference
## Prediction admit border notadmit
##   admit       29      2        0
##   border       1     25        0
##   notadmit     0      2       26
```

Expected:

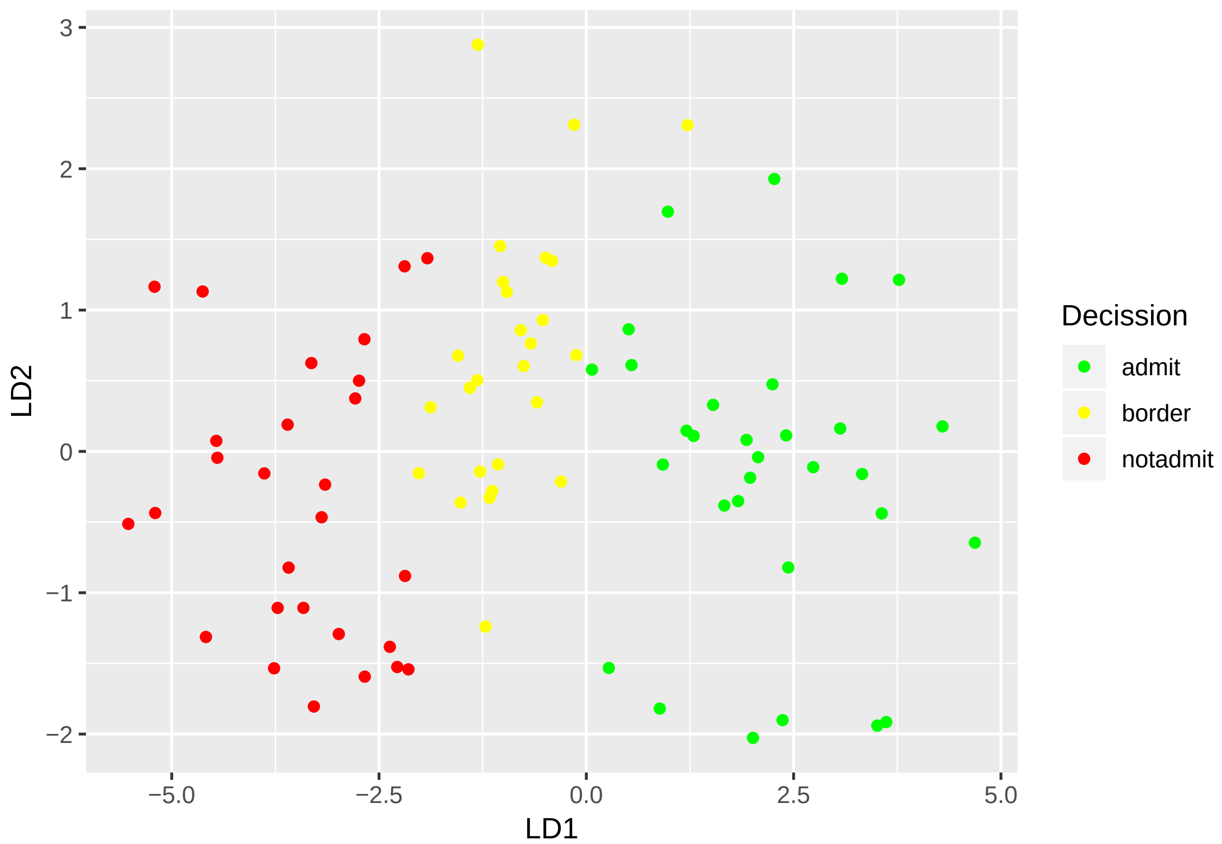
```
##           Reference
## Prediction     admit    border notadmit
##   admit    10.941176 10.576471 9.482353
##   border    9.176471  8.870588 7.952941
##   notadmit  9.882353  9.552941 8.564706
```

Residuals:

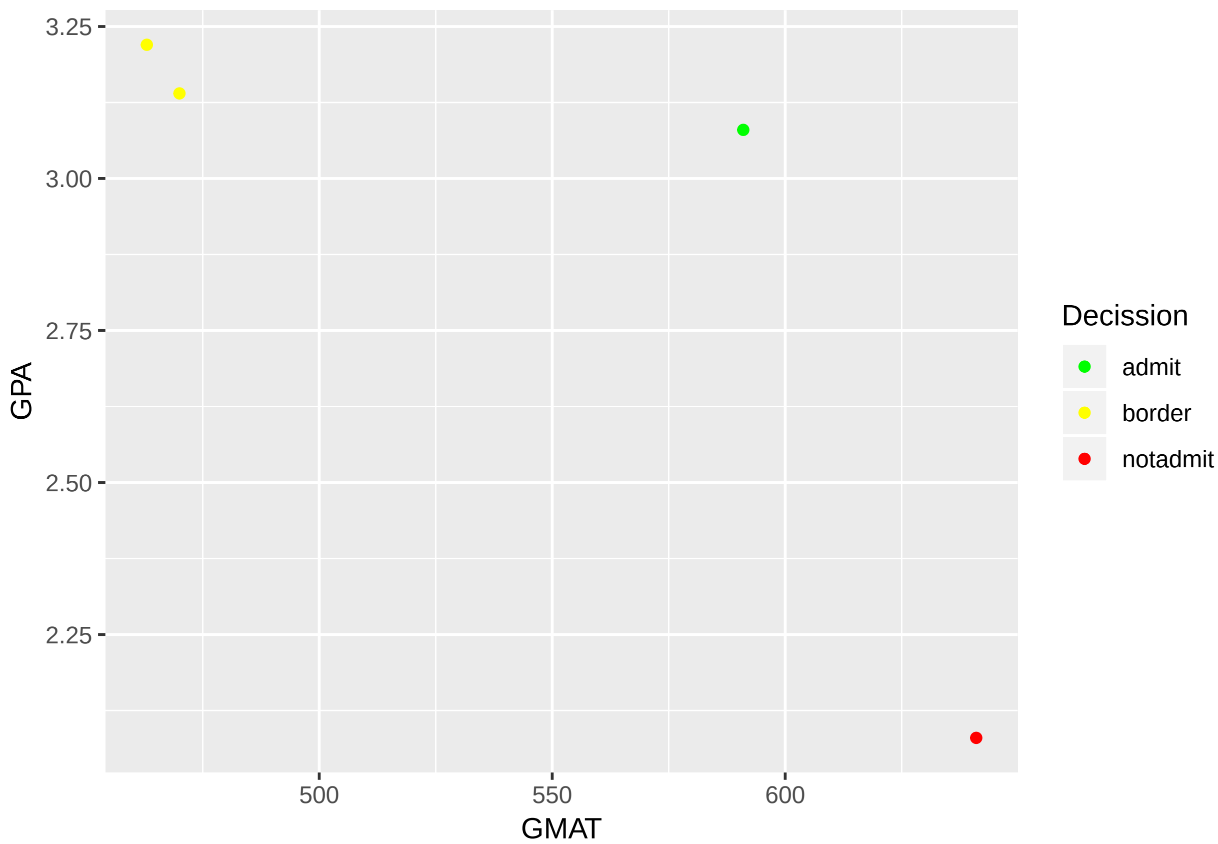
```
##           Reference
## Prediction     admit    border  notadmit
##   admit     5.459557 -2.637171 -3.079343
##   border   -2.699156  5.415547 -2.820096
##   notadmit -3.143621 -2.443698  5.957623
```

Standardized residuals:

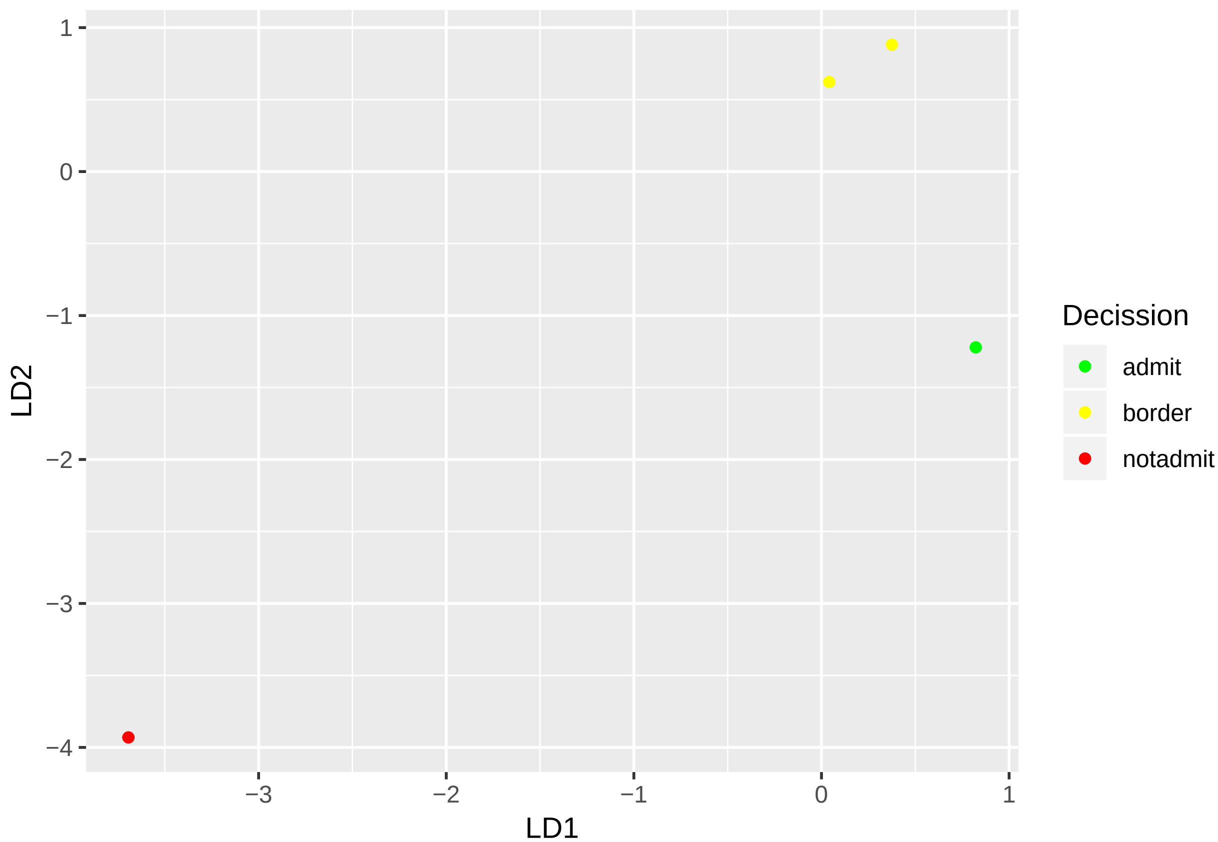
```
##           Reference
## Prediction     admit    border  notadmit
##   admit     8.515265 -4.076301 -4.637182
##   border   -4.027538  8.008316 -4.062850
##   notadmit -4.772329 -3.676504  8.732298
```



4. Predict students with GPA and GMAT score as below.



3.25
3.00
2.75
2.50
2.25
GPA
500
550
600
GMAT
Decission
admit
border
notadmit



GPA	GMAT	LD1	LD2	Decission
3.14	470	0.0410990	0.6216148	border
3.08	591	0.8222113	-1.2211957	admit
2.08	641	-3.6938624	-3.9305473	notadmit
3.22	463	0.3756372	0.8810811	border

Compare differences of the result from problem 1.

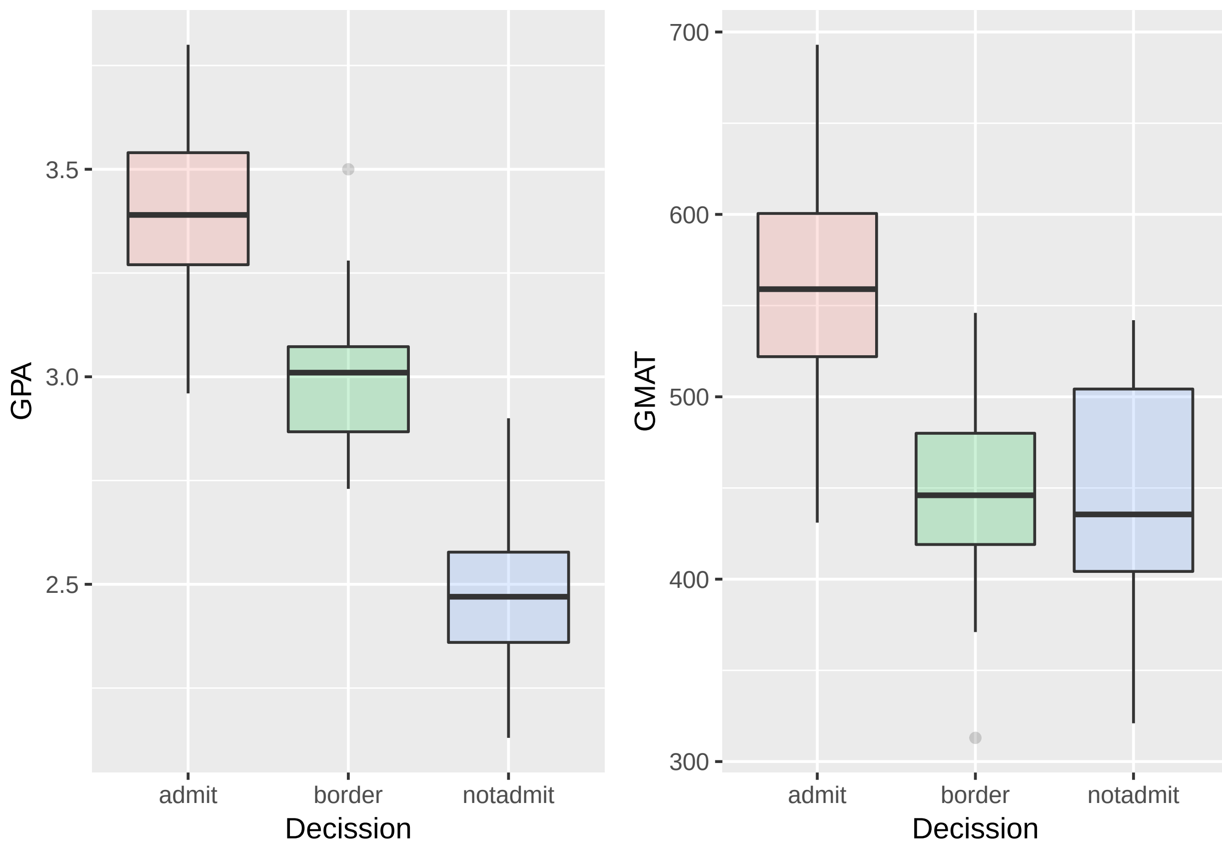
Both model predicted the same decission but the second model has an improved accurancy over the first first model.

Problem 3

Explain what is Quadratic Discriminant Analysis (QDA), and use QDA to train the model, discuss if this project can be done better by QDA, why or why not.

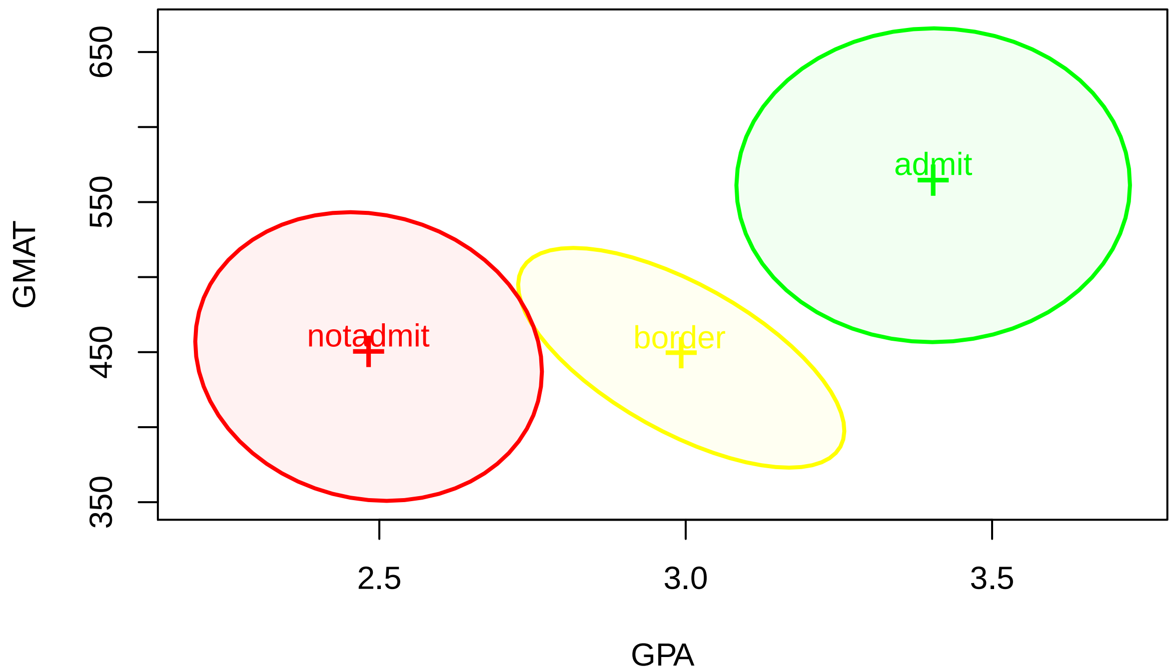
Quadratic discriminant analysis is a common tool for classification. QDA is used to determine which variables discriminate between two or more naturally occurring groups. QDA is closely related to LDA, where it assumes that the observations from each class of Y are drawn from a Gaussian distribution but assumes that each class has its own covariance matrix (i.e not identical). This project can not be done with QDA because the training set is not large enough.

Checking the Assumption of Equal Variance



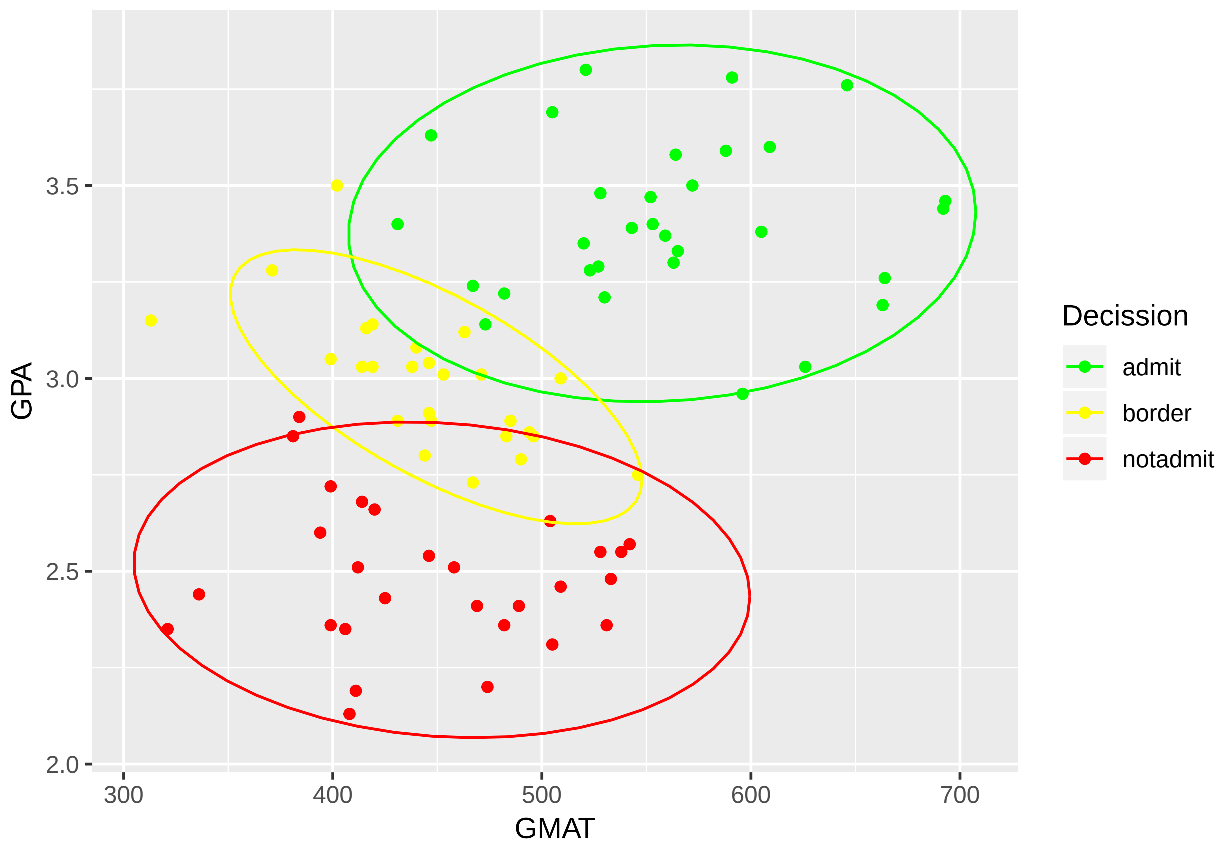
The two diferent boxplots show us that the length of each plot clearly differs. This is an indication for non-equal variances.

Checking the Assumption of Equal Covariance Ellipse




```
##
##  Bartlett test of homogeneity of variances
##
## data:  GPA by Decission
## Bartlett's K-squared = 1.0592, df = 2, p-value = 0.5888
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  GMAT by Decission
## Bartlett's K-squared = 3.4191, df = 2, p-value = 0.1809
```



From this scatterplot, we can clearly see that the variance for the admit and notadmit group is much wider than the variance from the border group. This is because the green and red points have a wider spread. The yellow points in contrast do not have as wide of a spread as the green and red points.

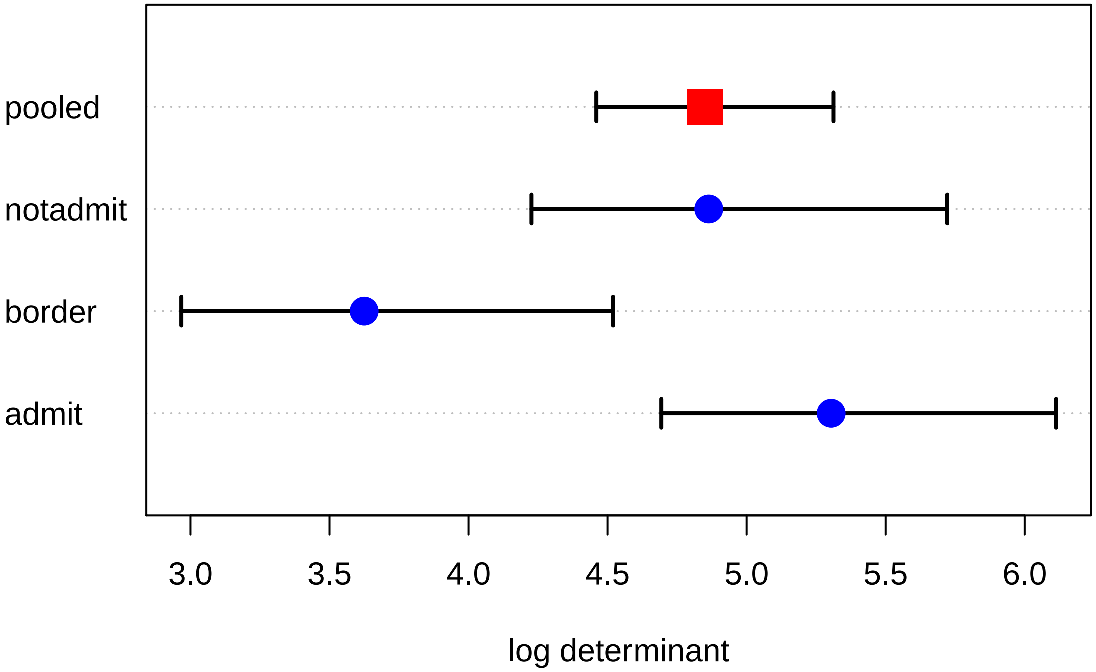
Use the BoxM test in order to check our assumption of homogeneity of variance-covariance matrices.

H_o = Covariance matrices of the outcome variable are equal across all groups

H_a = Covariance matrices of the outcome variable are different for at least one group

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  train.admission[, c(1:2)]
## Chi-Sq (approx.) = 16.074, df = 6, p-value = 0.01336
```

We reject the null hypothesis and conclude that we covariance matrices of the outcome variable for at least one group. The plot below gives information of how the groups differ in the components that go into Box's M test.



This plot confirms the visualizations that we have ellipses of different sizes and therefore, no equal variance-covariance matrices.

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value  Pr(>F)
## group  2   2.402 0.09688 .
##       82
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2   0.551 0.5785
##       82
```

```
# Fit the model
model <- qda(Decission~., data = train.admission)
model
```

```
## Call:
## qda(Decission ~ ., data = train.admission)
##
## Prior probabilities of groups:
##     admit    border  notadmit
## 0.3647059 0.3058824 0.3294118
```

```
##
## Group means:
##                GPA     GMAT
## admit    3.403871 561.2258
## border   2.992692 446.2308
## notadmit 2.482500 447.0714
```