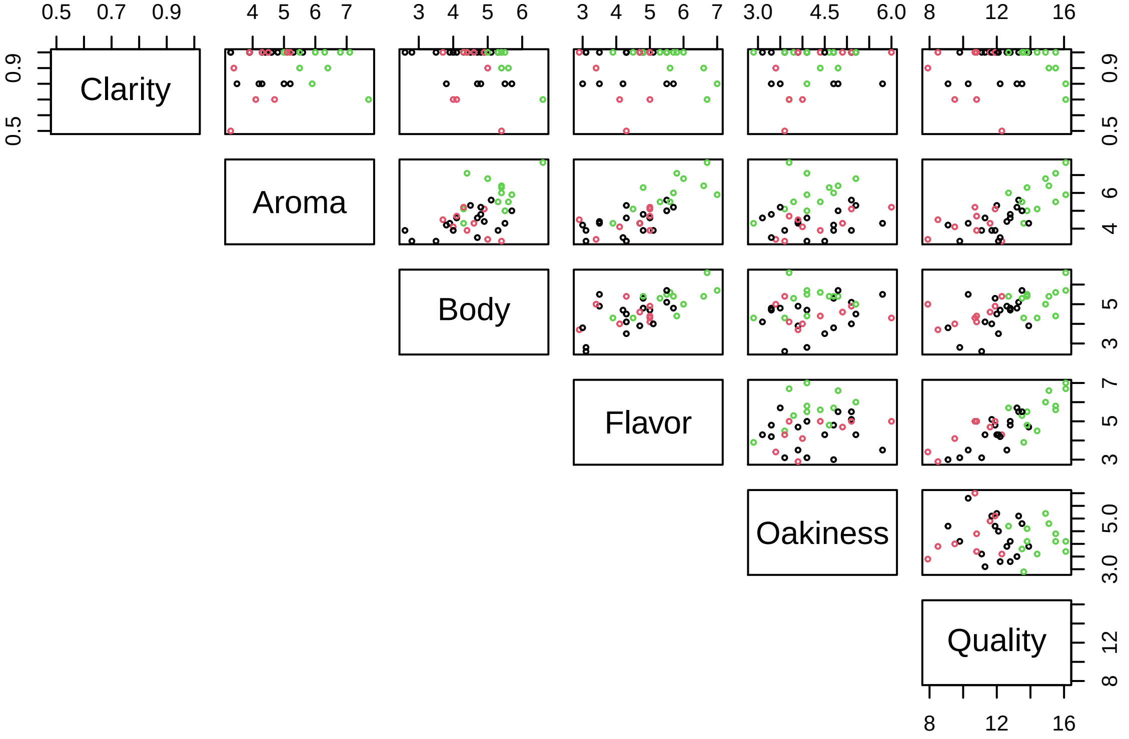
MiniProj_2

Wenxiong Lu

Q1

(a) The are 17 observations from region 1, 9 obs. from region 2 and 12 obs. from region 3. It seems obs. from region 3 have highest quality and obs. from region 2 and 1 are mixed.



(b): Yes. Quality is an approriate variable

(c):

Variable Name	P - Value	95%Confidence Interval
Clarity	0.865	(-5.105130, 6.043584)
Aroma	6.87e-07	(0.8850212 1.787982)
Body	0.000361	(1.159604 1.984177)
Flavor	3.68e-09	(1.159604 1.984177)
Oakiness	0.779	(-1.066083 0.805353)
Region 1	2e-16	(11.330893 12.6220486)
Region 2	0.00757	(-2.629298 -0.4347544)
Region 3	7.01e-06	(1.603271 3.6104546)
Region(overall)		

Comments: As the results shown above, ‘Clariy’ and ‘Oakiness’ are not significantly associated to Quality per 95% confidence level. The rest of variables are associated to Quality.

(d): Without considering interaction term, the remaining predictors are ‘Flavor’ and ‘Region’.(The interaction terms will be tested in question e) We can reject H0 for the F test of these parameters. Firstly, fit a full model with all the parameters, then we can find out that Clarity has highest p-value ‘0.990736’. After removing Clarity, the next highest will be removed. The ‘removing sequence’ is Clarity(0.990736), Body(0.746249), Aroma(0.70489), Oakiness(0.128060).

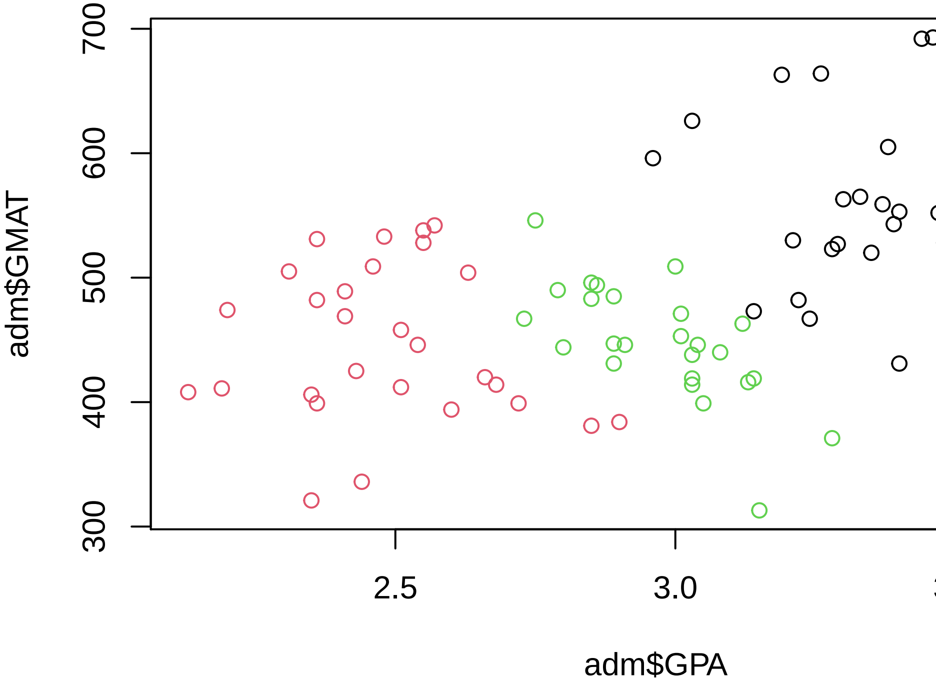
(e) Firstly, fit a full model (including all the posible interaction terms). Then remove terms with highest p-value and make anova for the updated model and find the next highest p-value and remove the predictor. By repeating this process we will remove the predictors in sequence: [Body:Flavor(interaction term between Body and Flavor), Aroma:Region, Flavor:Region, Aroma:Flavor:Region, Body:Flavor:Region, Aroma:Flavor, Aroma:Body:Region, Aroma:Body:Flavor:Region, Aroma:Body:Flavor, Aroma:Body, Body:Region]. The remaing predictors are Aroma, Body, Flavor, Region. Then, check the F test for multiple linear model and we can see that the p-value of Aroma and Body changed from significant to insignificant. This is due to Region is a qualitative data, both Aroma and Body have significant interaction to Region 1 and Region 2. The currently reasonably good model is: #####Quality = x0+Aroma*x1+Body*x2+Flavor*x3+Region*x4.

```
## Analysis of Variance Table
##
## Response: Quality
##           Df Sum Sq Mean Sq F value    Pr(>F)    
## Aroma      1 77.442  77.442 91.2025 6.909e-11 ***
## Body       1  5.703   5.703  6.7163   0.01428 *  
## Flavor     1 18.878  18.878 22.2329 4.539e-05 ***
## Region     2 25.593  12.797 15.0706 2.445e-05 ***
## Residuals 32 27.172   0.849                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Call:
## lm(formula = Quality ~ Aroma + Body + Flavor + Region, data = wine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.98279 -0.59142  0.02005  0.55790  1.87722 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  7.00195    1.11063   6.304 4.51e-07 ***
## Aroma       -0.01643    0.24448  -0.067 0.946855    
## Body         0.05253    0.24509   0.214 0.831664    
## Flavor       1.10027    0.24136   4.559 7.13e-05 ***
## Region2     -1.53977    0.38117  -4.040 0.000313 ***
## Region3      1.22420    0.47800   2.561 0.015352 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9215 on 32 degrees of freedom
## Multiple R-squared:  0.8245, Adjusted R-squared:  0.797 
## F-statistic: 30.06 on 5 and 32 DF,  p-value: 3.341e-11
```

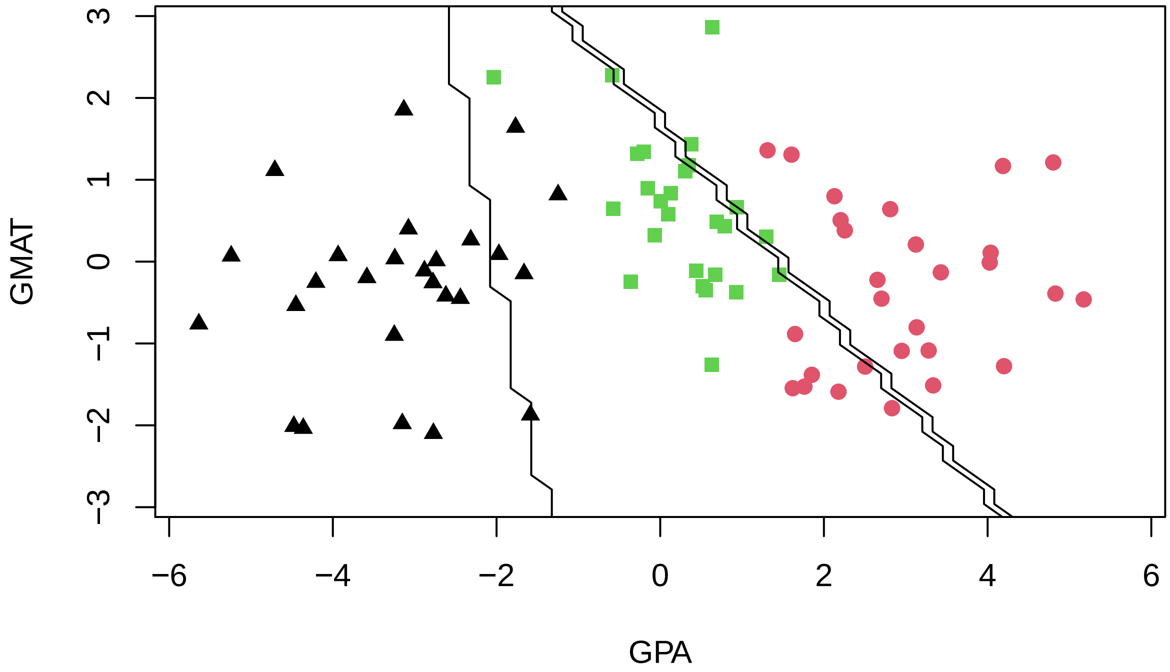
(f) By fiting a linear model for Aroma = Region*Body, we can find out the exact interaction term in (e) is of 1.(Region 1, Body, Region2:Body) and fit another model Body = Aroma*Region we can see interaction term is of 2.(Aroma, Region2, Aroma:Region2). This indicates that there are interaction terms between Region1 and Aroma, Region and Body. So we should consider put back the interaction term between Region:Aroma and Region:Body.

The final model: Quality = x0 + Aroma x1 + Body x2 + Flavor x3 + Region x4 + Aroma:Region x5 + Body:Region x6



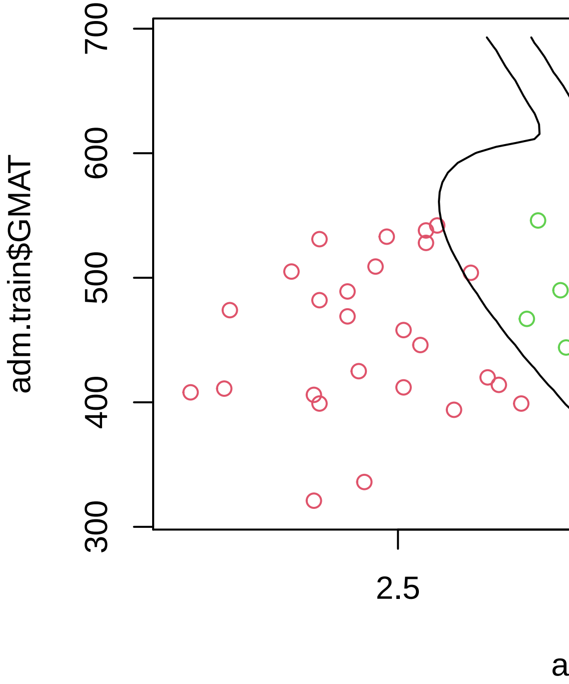
##### 2(a). It is easy to see that group is related to GPA and GMAT. Group might be an important predictor. GPA and GMAT are related to each differently in different group.

```
## Call:
## lda(Group ~ GPA + GMAT, data = adm.train)
##
## Prior probabilities of groups:
##     1     2     3
## 0.325 0.350 0.325
##
## Group means:
##        GPA     GMAT
## 1 3.431538 569.8077
## 2 2.482500 447.0714
## 3 2.992692 446.2308
##
## Coefficients of linear discriminants:
##               LD1         LD2
## GPA  -5.300511724  1.91775603
## GMAT -0.009125023 -0.01438851
##
## Proportion of trace:
##   LD1   LD2
## 0.969 0.031
```



```
##
##      1  2  3
##   1 24  0  2
##   2  0 26  2
##   3  0  0 26

## [1] "confusion matrix for training data"

## [1] 0.05

## [1] "confusion matrix for test data"

##
##     1 2 3
##   1 2 0 3
##   2 0 0 0
##   3 0 0 0

## [1] 0.6
```

(b) The decision boundary seems sensible. The misclassification rate:p(training data) = 0.05, p(test data) = 0.6 There is exist overfiting problem in this model.



```
##
##       1  2  3
##    1 26  0  0
##    2  0 27  1
##    3  1  0 25
```

```
## [1] "confusion matrix for training data"
```

```
## [1] 0.05
```

```
## [1] "confusion matrix for test data"
```

```
##
##      1 2 3
##    1 4 0 1
##    2 0 0 0
##    3 0 0 0
```

```
## [1] 0.2
```

(c) The decision boundary seems sensible. The misclassification rate:p(training data) = 0.025, p(test data) = 0.2 The model predicts the test data well.

(d) Ingeneral, QDA performs better than LDA in this case.So QDA is recommend.

3 It seems BloodyPressure, Insulin, age, Glucose, are strongly related to Outcome.