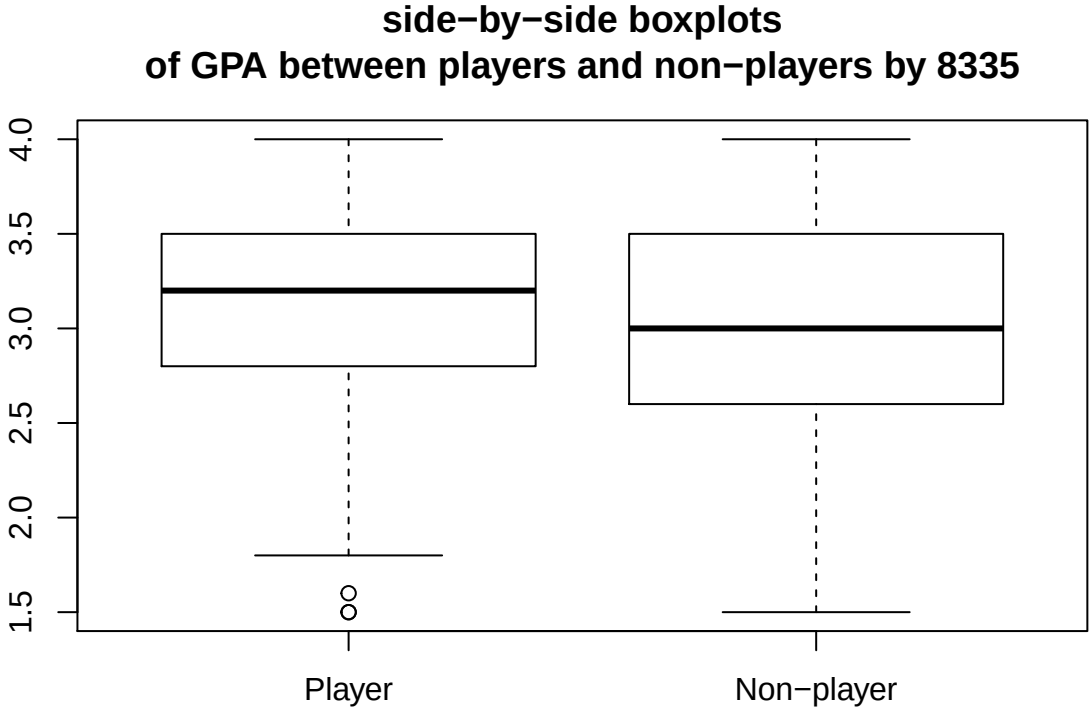
STA303 Assignment 2

Haoda Li 1003918335

Solutions

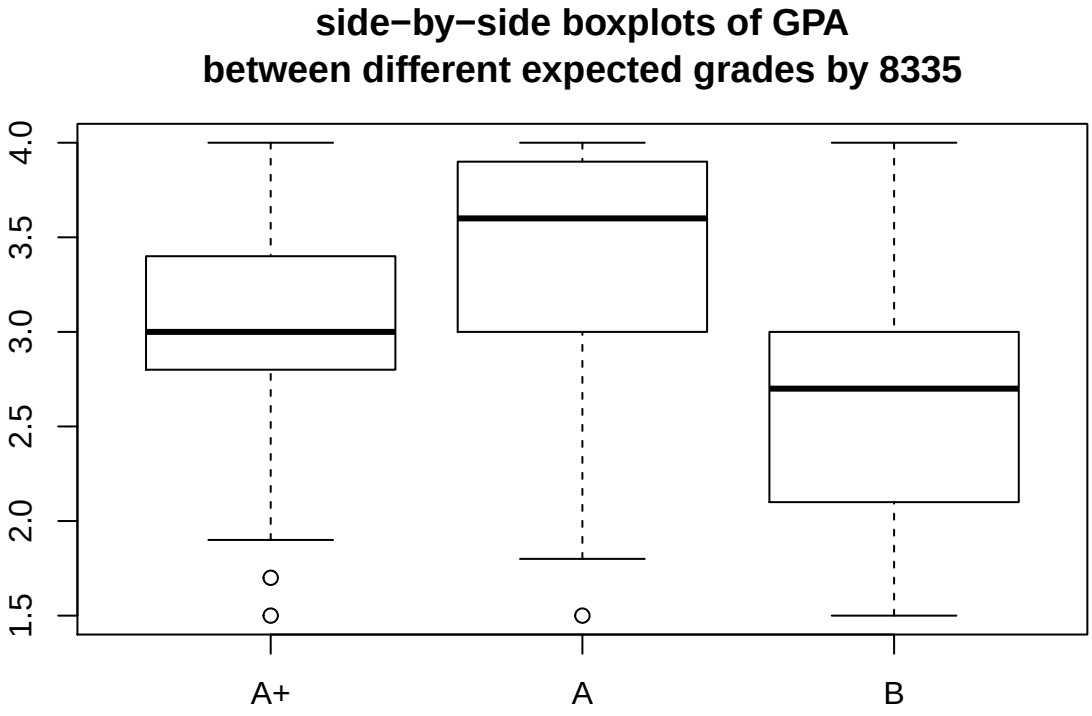
Question 1

i.



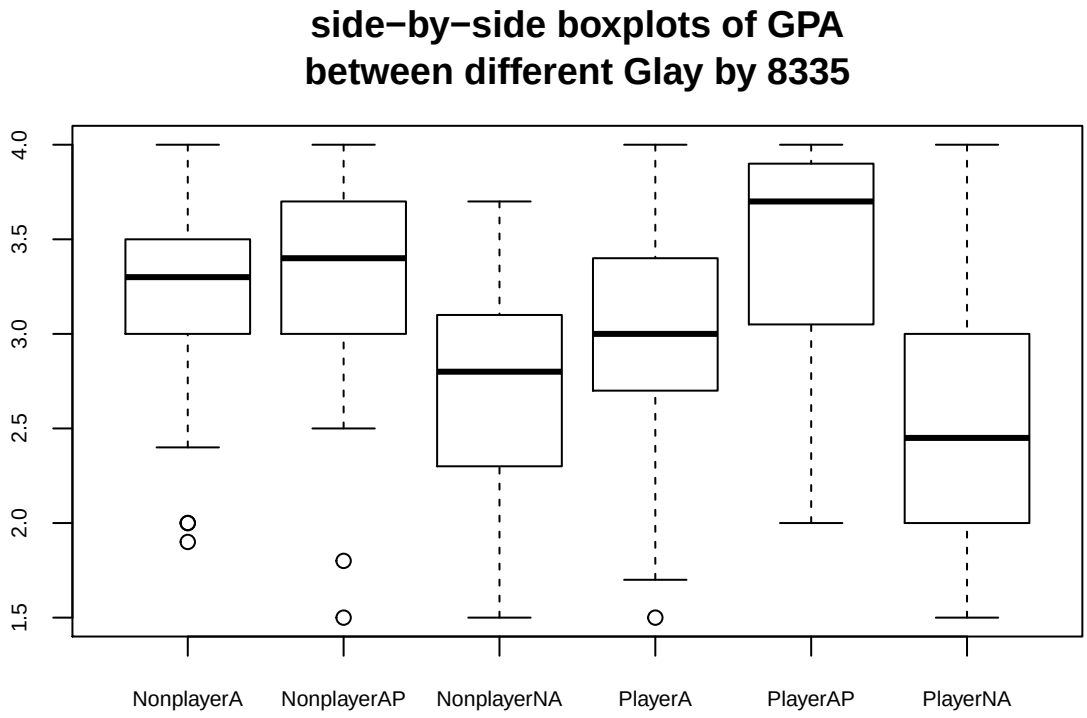
The difference of median, third quartile, maximum of GPA between players and non-players are very small, while the first quartile and minimum have some difference.

ii.



There is a significant difference in mean, first quartile, third quartile, minimum, and maximum of GPA among the three groups.

iii.



There is a significant difference in mean, first quartile, third quartile, minimum, and maximum of GPA among the three groups.

Question 2

```
##
##  Welch Two Sample t-test
##
## data:  GPA by Player
## t = 1.1831, df = 187.34, p-value = 0.2383
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.05394441  0.21561458
## sample estimates:
## mean in group 0 mean in group 1
##        3.082524        3.001689
```

Two sample t-test
Null hypothesis: $H_0 : \mu_p - \mu_n = 0$ where μ_p is the mean GPA of players, μ_n is the mean GPA of non-players.

Test statistic: 1.1831

p-value: 0.2383

By the result of a pooled two sample t-test, since the p-value is $0.23 > 0.05$, we cannot reject the null hypothesis. Therefore, there is no evidence suggesting that there is a difference in means between Players and Non-players.

Question 3

One way ANOVA

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Grade         2  34.87  17.434   59.84 <2e-16 ***
## Residuals   396 115.37   0.291
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null hypothesis: $H_0 : \mu_{A+} = \mu_A = \mu_B$ where μ_{A+}, μ_A, μ_B are mean GPA of students with expected grades $A+, A, B$.

Alternative hypothesis: H_a: at least one pair of μ_{A+}, μ_A, μ_B does not equal.

Test statistic: 59.84

p-value: ≈ 0.0

By the result of one way ANOVA, since the p-value is approximately 0, we can reject the null hypothesis. Therefore, there is some evidence that there is a difference in mean between students with different expected grades.

Pairwise comparisons

Since there are only 3 levels of expected grades, I'll use Bonferroni's Method for pairwise comparisons.

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  GPA and Grade
##
##     A       A+
## A+  1.8e-07 -
## B   2.6e-11 < 2e-16
##
## P value adjustment method: bonferroni
```

By the result of pair wise t-test, for each pair of comparisons, the p-value is approximately 0. Therefore, there is some evidence that the GPA differs for each pair among all levels of expected grades.

Question 4

One way ANOVA

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Glay          5  37.15   7.431   25.82 <2e-16 ***
## Residuals   393 113.08   0.288
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null hypothesis: H_0: the mean GPA is equal for all six categories of students classfied by the combination of their player status and expected grade.

Alternative hypothesis: H_a: at least one pair of categories among the six categories of students classfied by the combination of their player status and expected grade does not equal.

Test statistic: 25.82

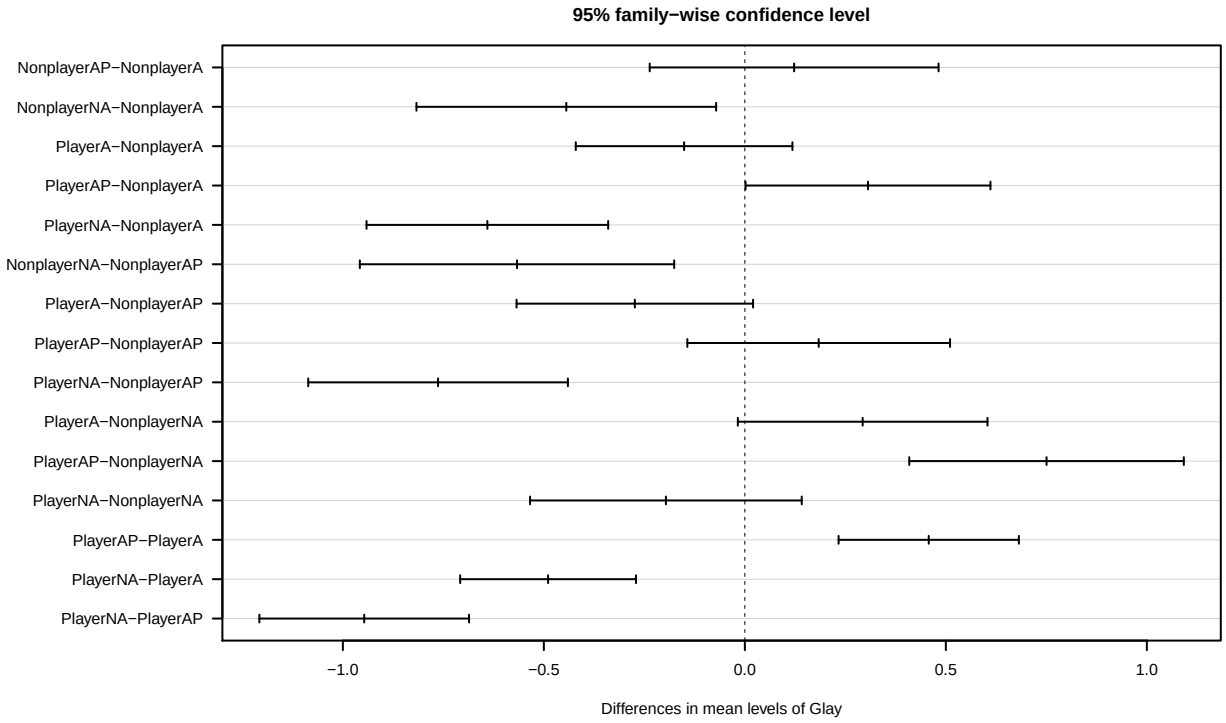
p-value: ≈ 0.0

By the result of one way ANOVA, since the p-value is approximately $0 \leq 0.05/3$, we can reject the null hypothesis. Therefore, there is some evidence that there is a difference in mean between students with different categories of students classfied by the combination of their player status and expected grade.

Pairwise comparisons

Since there are 6 levels of expected grades, I'll use Tukey's Method for pairwise comparisons.

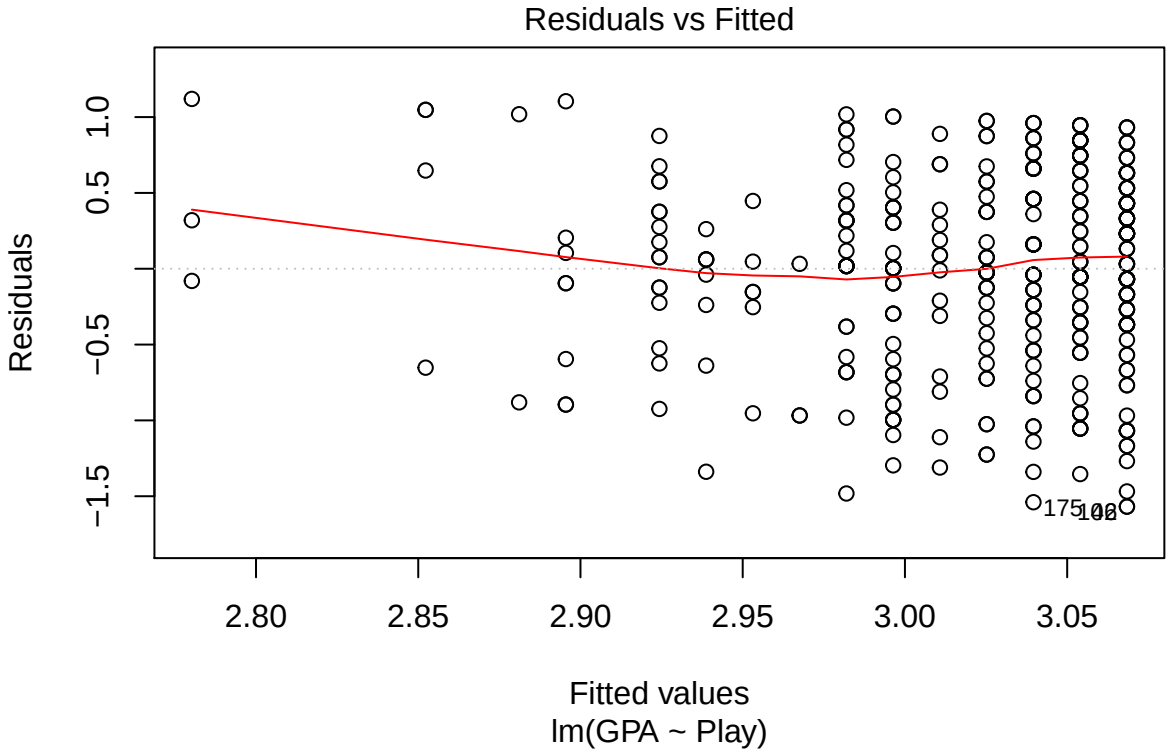
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = GPA ~ Glay, data = student)
##
## $Glay
##                                diff          lwr         upr     p adj
## NonplayerAP-NonplayerA   0.1226164 -0.236652579  0.48188540 0.9249364
## NonplayerNA-NonplayerA  -0.4441548 -0.816898923 -0.07141058 0.0092179
## PlayerA-NonplayerA      -0.1510952 -0.420523778  0.11833332 0.5950421
## PlayerAP-NonplayerA      0.3063342  0.001730383  0.61093798 0.0477882
## PlayerNA-NonplayerA     -0.6405149 -0.941076726 -0.33995308 0.0000000
## NonplayerNA-NonplayerAP -0.5667712 -0.957785463 -0.17575686 0.0005766
## PlayerA-NonplayerAP     -0.2737116 -0.567898161  0.02047488 0.0848409
## PlayerAP-NonplayerAP     0.1837178 -0.142989193  0.51042474 0.5921294
## PlayerNA-NonplayerAP    -0.7631313 -1.086073067 -0.44018956 0.0000000
## PlayerA-NonplayerNA      0.2930595 -0.017439629  0.60355867 0.0768963
## PlayerAP-NonplayerNA     0.7504889  0.409019383  1.09195849 0.0000000
## PlayerNA-NonplayerNA    -0.1963602 -0.534229046  0.14150874 0.5562541
## PlayerAP-PlayerA         0.4574294  0.233253189  0.68160564 0.0000002
## PlayerNA-PlayerA        -0.4894197 -0.708072168 -0.27076718 0.0000000
## PlayerNA-PlayerAP       -0.9468491 -1.207618423 -0.68607975 0.0000000
```

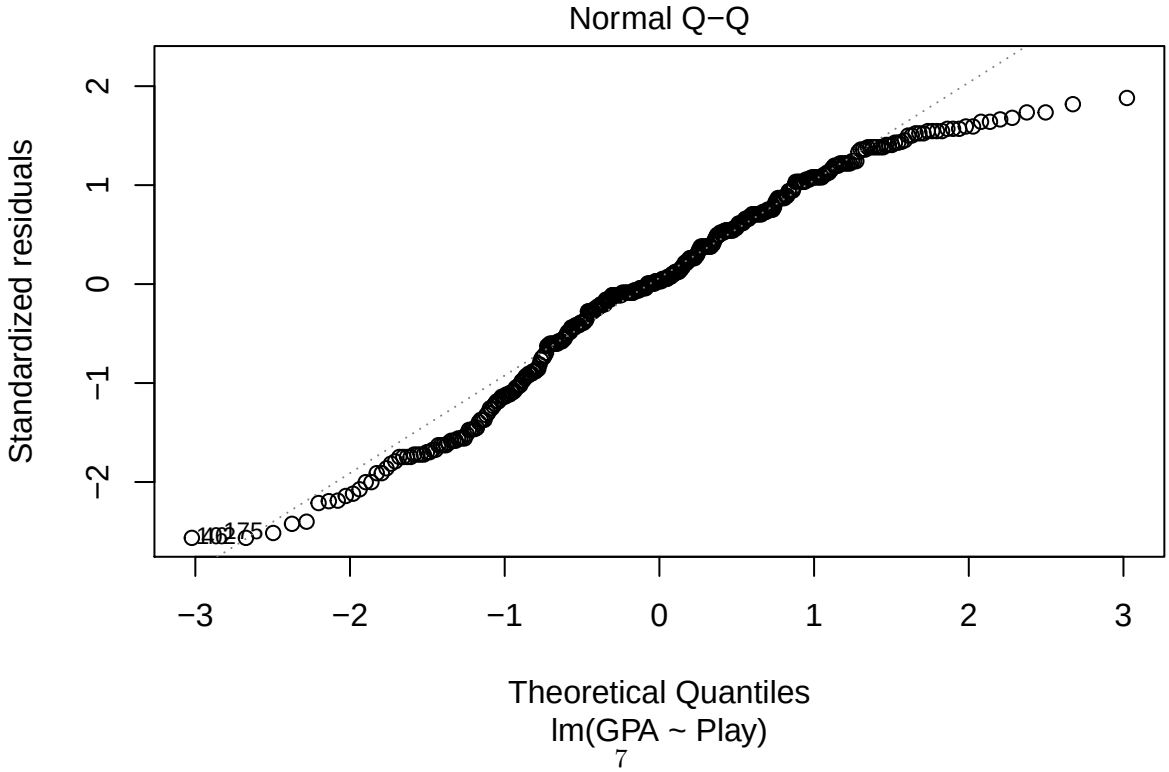


By the output of Tukey's HSD method, the pair of catories that have different mean GPA are:

- NonplayerNA-NonplayerA
- PlayerAP-NonplayerA
- PlayerNA-NonplayerA
- NonplayerNA-NonplayerAP
- PlayerNA-NonplayerAP
- PlayerAP-NonplayerNA
- PlayerAP-PlayerA
- PlayerNA-PlayerA
- PlayerNA-PlayerAP

Question 5


Residuals vs Fitted
Residuals
1.0
0.5
−0.5
−1.5
2.80
2.85
2.90
2.95
3.00
3.05
Fitted values
lm(GPA ~ Play)


Normal Q−Q
Standardized residuals
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(GPA ~ Play)

```
##
##  Bartlett test of homogeneity of variances
##
## data:  GPA by Play
## Bartlett's K-squared = 5.8108, df = 14, p-value = 0.971
```

From the residuals vs. fitted plot, we can observe that the points spread out. Also, the Bartlett test has a extreme large p-value (0.971). The assumption of constant variance is violated.

From the Normal Q-Q plot, we can observe that the plot is heavy-tailed. The assumeption of normality of errors is violated.

we should not be concerned that the data contained different numbers of students in the three grade levels. It is not a part of our assumptions and the group size will not influence the test statistics significantly.

Question 6

a) $Y_i = \beta_0 + \beta_1 X_{A,i} + \beta_2 X_{B,i} + \beta_3 X_{p,i} + \beta_4 X_{A,i} X_{p,i} + \beta_5 X_{B,i} X_{p,i}$ where Y_i is the GPA of ith student; $X_{A,i}, X_{B,i}$ are the indicators that the ith student's expected grade is A, B, respectively; $X_{p,i}$ is the indicator that the ith student is a player.

b) No, the total number of predictor variables are $(2-1)+(3-1)+(2-1)(3-1) = 5 \neq 6$ since there are 2 levels of Player status and 3 levels of expected grades.

c) The F-test will be significant. From the result of Q4, we have evidence that there is difference among pairs such as PlayerNA-NonplayerAP, PlayerAP-NonplayerNA, PlayerAP-NonplayerA, PlayerNA-NonplayerA. These results shows that there are some interactions between two variables in explaining GPA.

Question 7

The mathematical equation is $Y_i = \beta_0 + \beta_1 X_{A,i} + \beta_2 X_{B,i} + \beta_3 X_{p,i}$ where Y_i is the GPA of ith student; $X_{A,i}, X_{B,i}$ are the indicators that the ith student's expected grade is A, B, respectively; $X_{p,i}$ is the number of hours the ith student spent playing video or computer games.

The new model treats Play as a continous variable rather than a categorical variable.

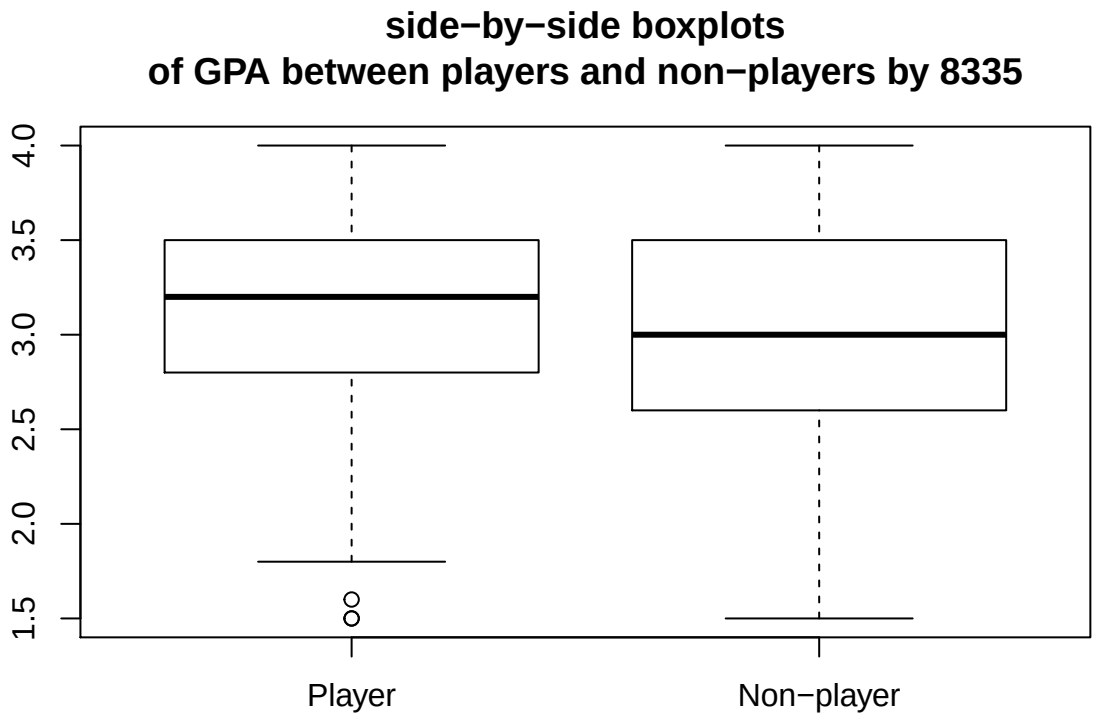
Question 8

Factor: the current status of student. Levels: part-time, full-time.

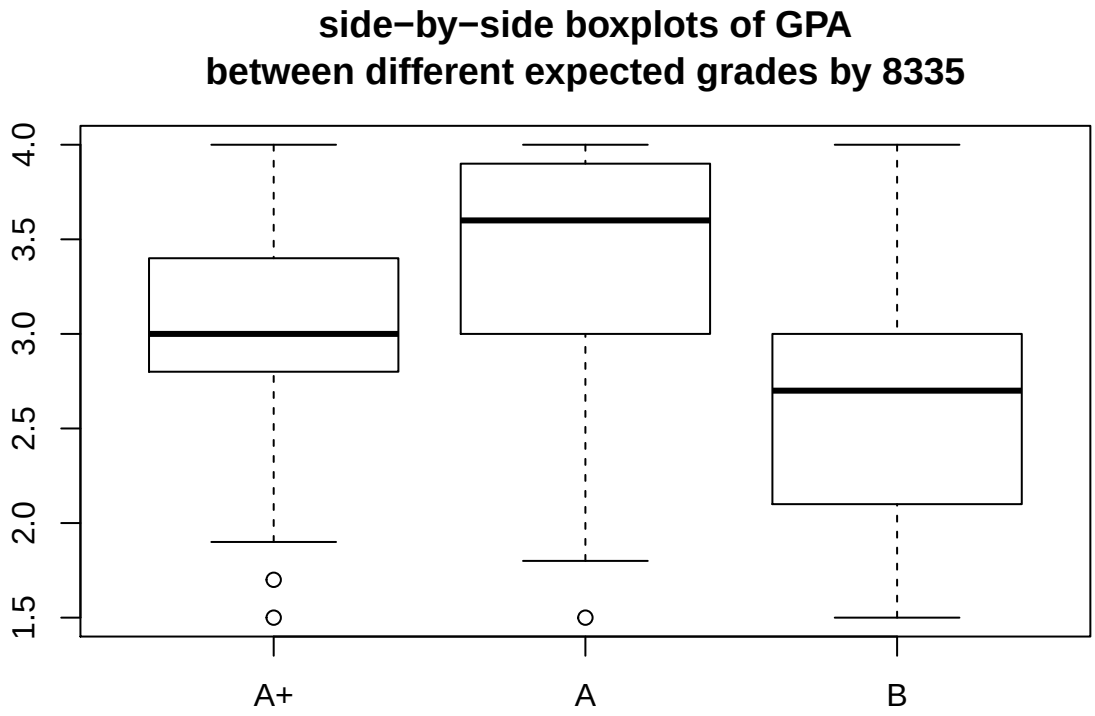
Factor: the student's background of English. Levels: native speaker, has over 5 years of experience, less than or equal to 5 years of experience.

Appendix

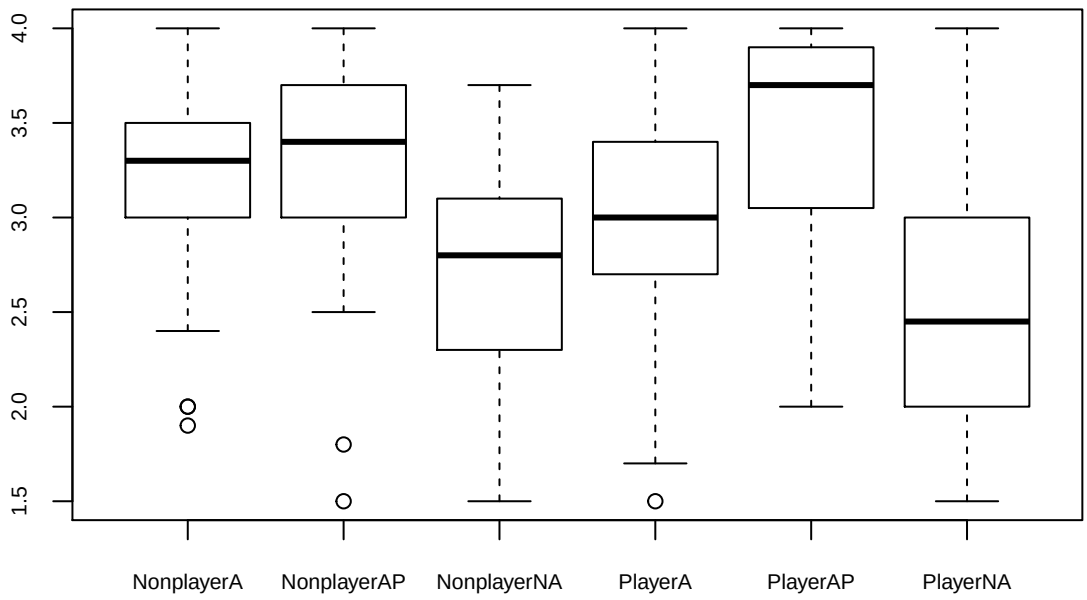
```
student <- read.csv('data2.csv')
GPA <- student$GPA
Grade <- student$Grade
Player <- as.integer(student$Play != 0)
Glay <- NULL
for (i in 1:399)
{ if (Player[i]==0 & Grade[i]=="B ")
{Glay[i]="NonplayerNA"}
else if (Player[i]==0 & Grade[i]=="A ")
{Glay[i]="NonplayerA"}
else if (Player[i]==0 & Grade[i]=="A+ ")
{Glay[i]="NonplayerAP"}
else if (Player[i]==1 & Grade[i]=="B " )
{Glay[i]="PlayerNA"}
else if (Player[i]==1 & Grade[i]=="A ")
{Glay[i]="PlayerA"}
else {Glay[i]="PlayerAP"}
}
Player=as.factor(Player)
Glay=as.factor(Glay)
boxplot(GPA~Player,data=student, names=c('Player', 'Non-player'),
        main="side-by-side boxplots\nof GPA between players and non-players by 8335")
```



```
boxplot(GPA~Grade, data=student, names=c('A+', 'A', 'B'),
        main="side-by-side boxplots of GPA\nbetween different expected grades by 8335")
```




```
par(cex.axis=0.7)
boxplot(GPA~Glay, data=student, main="side-by-side boxplots of GPA\nbetween different Glay by 8335")
```




```
t.test(GPA~Player, data=student)
```

```
##
##  Welch Two Sample t-test
##
## data:  GPA by Player
## t = 1.1831, df = 187.34, p-value = 0.2383
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.05394441  0.21561458
## sample estimates:
## mean in group 0 mean in group 1
##        3.082524        3.001689
```

```
gg <- aov(GPA~Grade, data=student)
summary(gg)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Grade         2  34.87  17.434   59.84 <2e-16 ***
## Residuals   396 115.37   0.291
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
pairwise.t.test(GPA, Grade, data=student, p.adj='bonf')
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
```

```
## data:  GPA and Grade
##
##     A       A+
## A+  1.8e-07 -
## B   2.6e-11 < 2e-16
##
## P value adjustment method: bonferroni
```

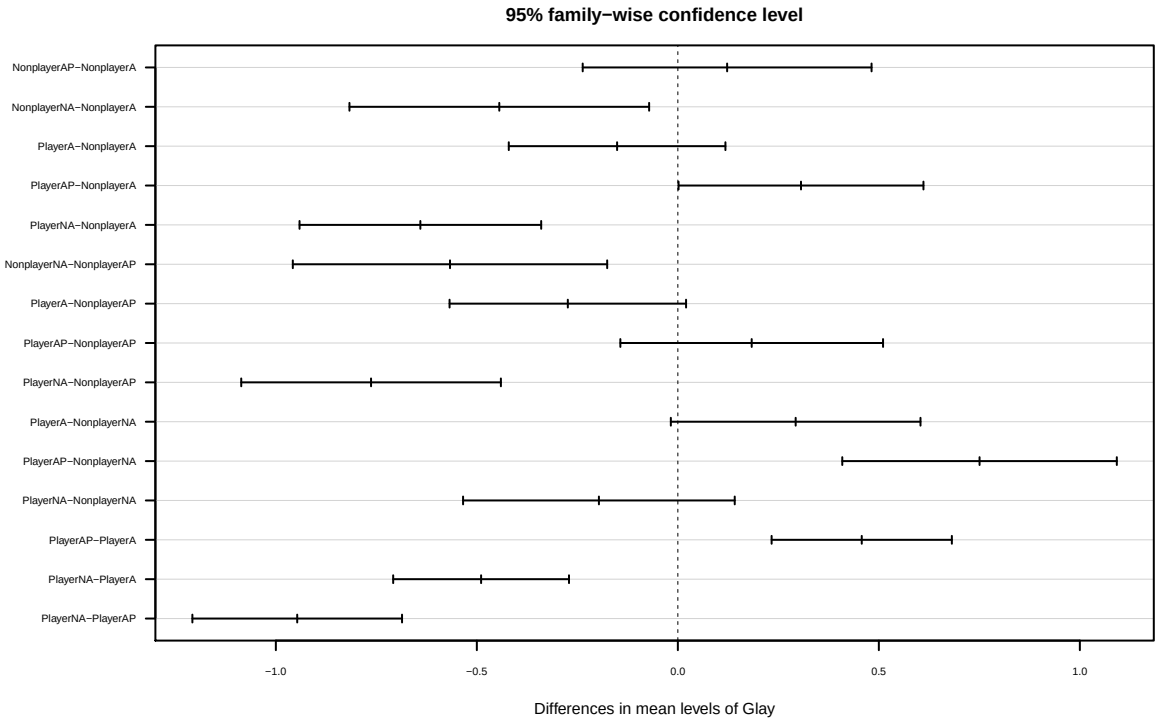
```
ggl <- aov(GPA~Glay, data=student)
summary(ggl)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Glay          5  37.15   7.431   25.82 <2e-16 ***
## Residuals   393 113.08   0.288
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

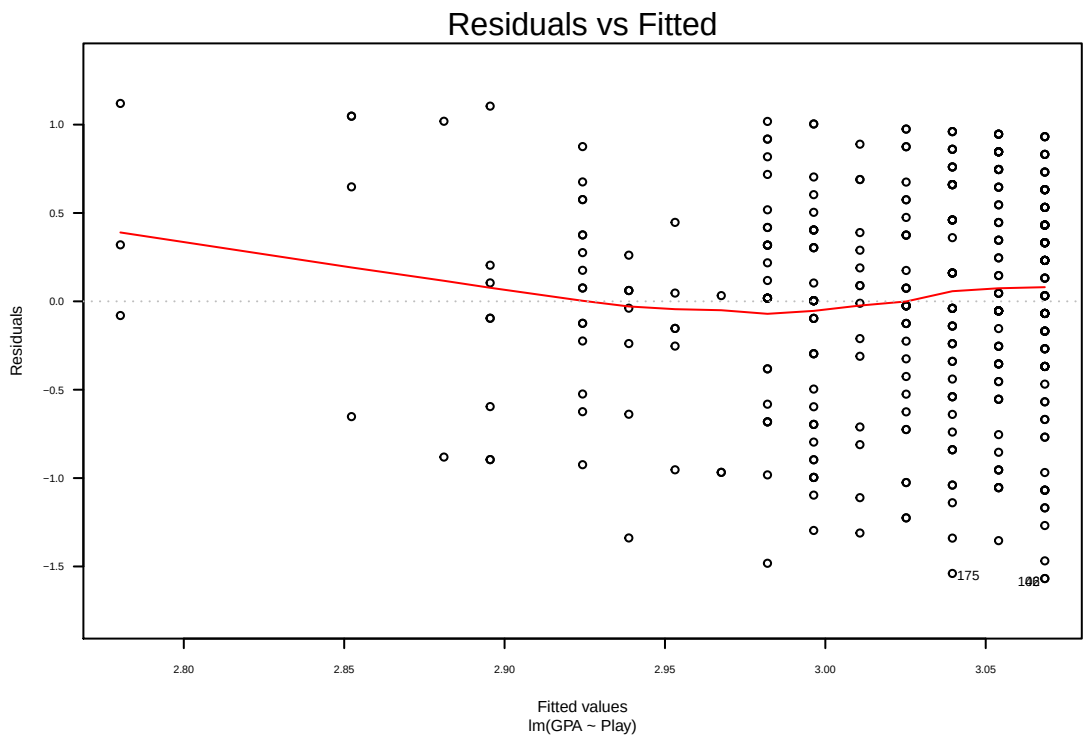
```
hsd <- TukeyHSD(ggl, "Glay")
hsd
```

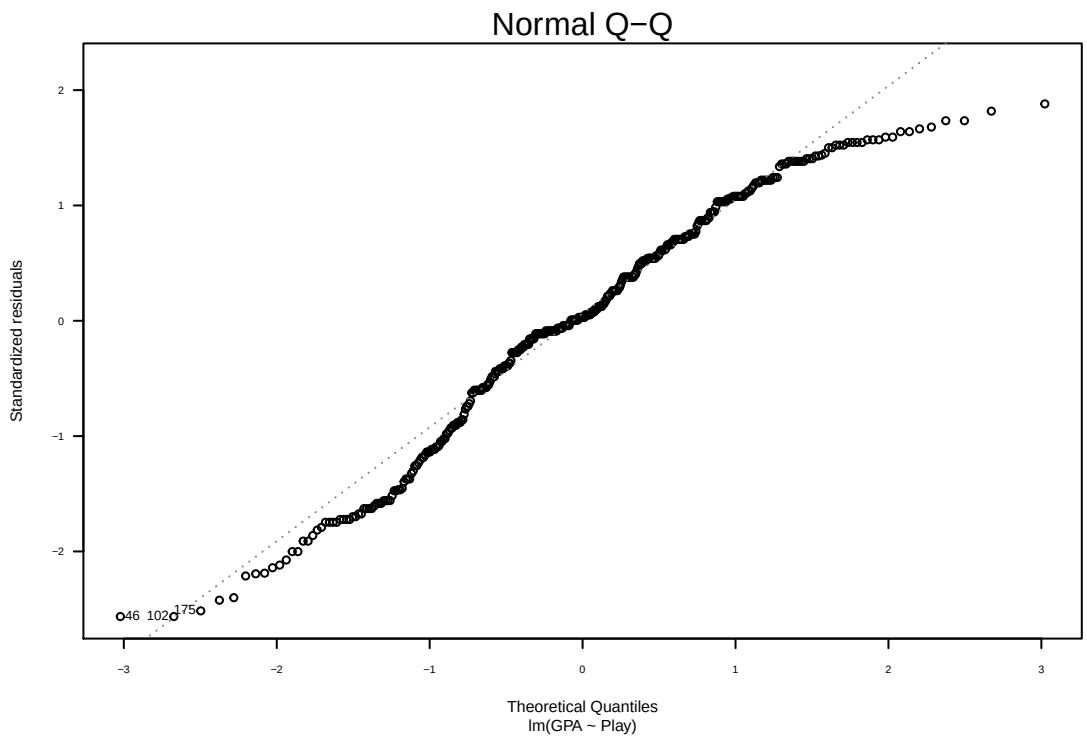
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = GPA ~ Glay, data = student)
##
## $Glay
##                                diff          lwr         upr     p adj
## NonplayerAP-NonplayerA    0.1226164 -0.236652579  0.48188540 0.9249364
## NonplayerNA-NonplayerA   -0.4441548 -0.816898923 -0.07141058 0.0092179
## PlayerA-NonplayerA       -0.1510952 -0.420523778  0.11833332 0.5950421
## PlayerAP-NonplayerA       0.3063342  0.001730383  0.61093798 0.0477882
## PlayerNA-NonplayerA      -0.6405149 -0.941076726 -0.33995308 0.0000000
## NonplayerNA-NonplayerAP  -0.5667712 -0.957785463 -0.17575686 0.0005766
## PlayerA-NonplayerAP      -0.2737116 -0.567898161  0.02047488 0.0848409
## PlayerAP-NonplayerAP      0.1837178 -0.142989193  0.51042474 0.5921294
## PlayerNA-NonplayerAP     -0.7631313 -1.086073067 -0.44018956 0.0000000
## PlayerA-NonplayerNA       0.2930595 -0.017439629  0.60355867 0.0768963
## PlayerAP-NonplayerNA      0.7504889  0.409019383  1.09195849 0.0000000
## PlayerNA-NonplayerNA     -0.1963602 -0.534229046  0.14150874 0.5562541
## PlayerAP-PlayerA          0.4574294  0.233253189  0.68160564 0.0000002
## PlayerNA-PlayerA         -0.4894197 -0.708072168 -0.27076718 0.0000000
## PlayerNA-PlayerAP        -0.9468491 -1.207618423 -0.68607975 0.0000000
```

```
par(mar=c(12,12,2,1), cex=0.5, las=1)
plot(hsd)
```




```
plot(lm(GPA~Play, data=student), which = 1:2)
```


Residuals vs Fitted
Residuals
1.0
0.5
0.0
−0.5
−1.0
−1.5
2.80
2.85
2.90
2.95
3.00
3.05
175
Fitted values
lm(GPA ~ Play)




```
bartlett.test(GPA~Play, data=student)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  GPA by Play
## Bartlett's K-squared = 5.8108, df = 14, p-value = 0.971
```