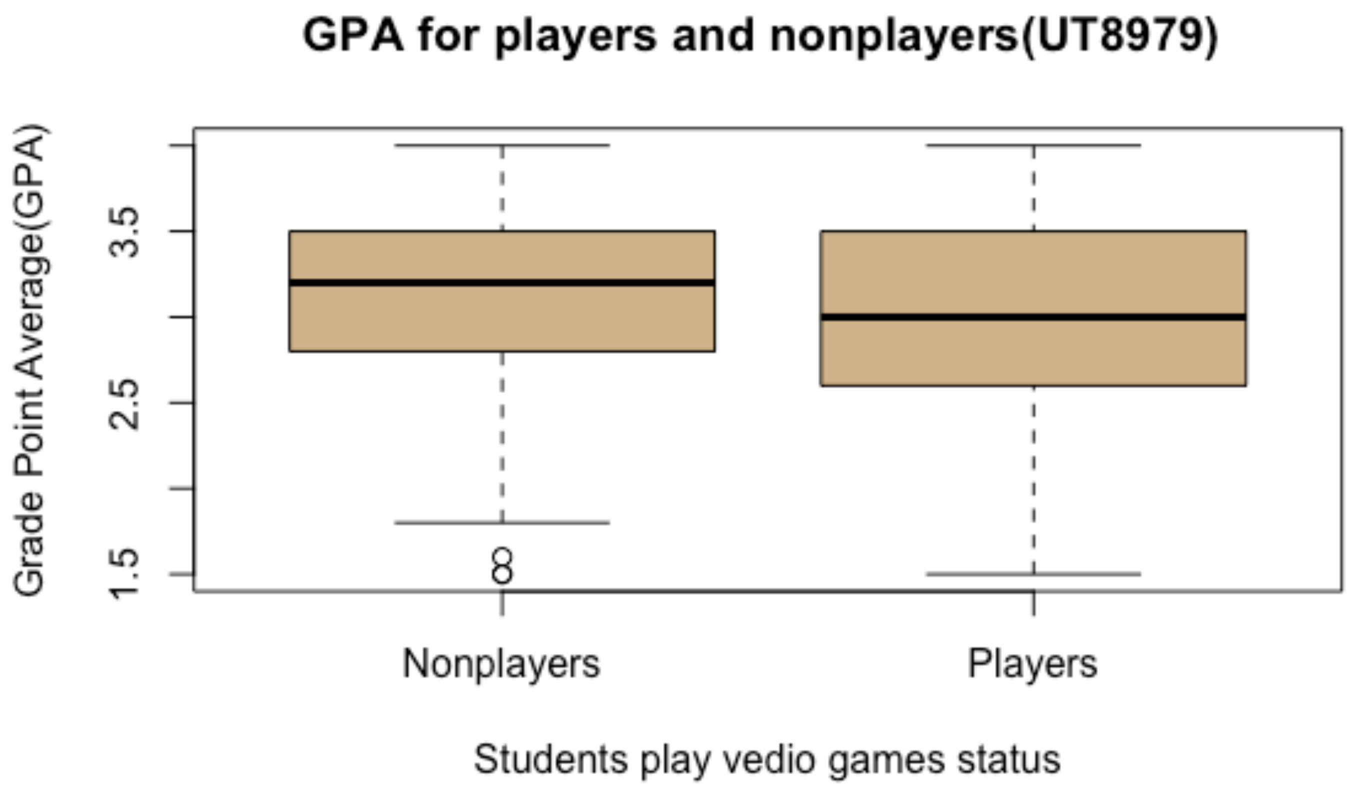
STA 303 Assignment 2

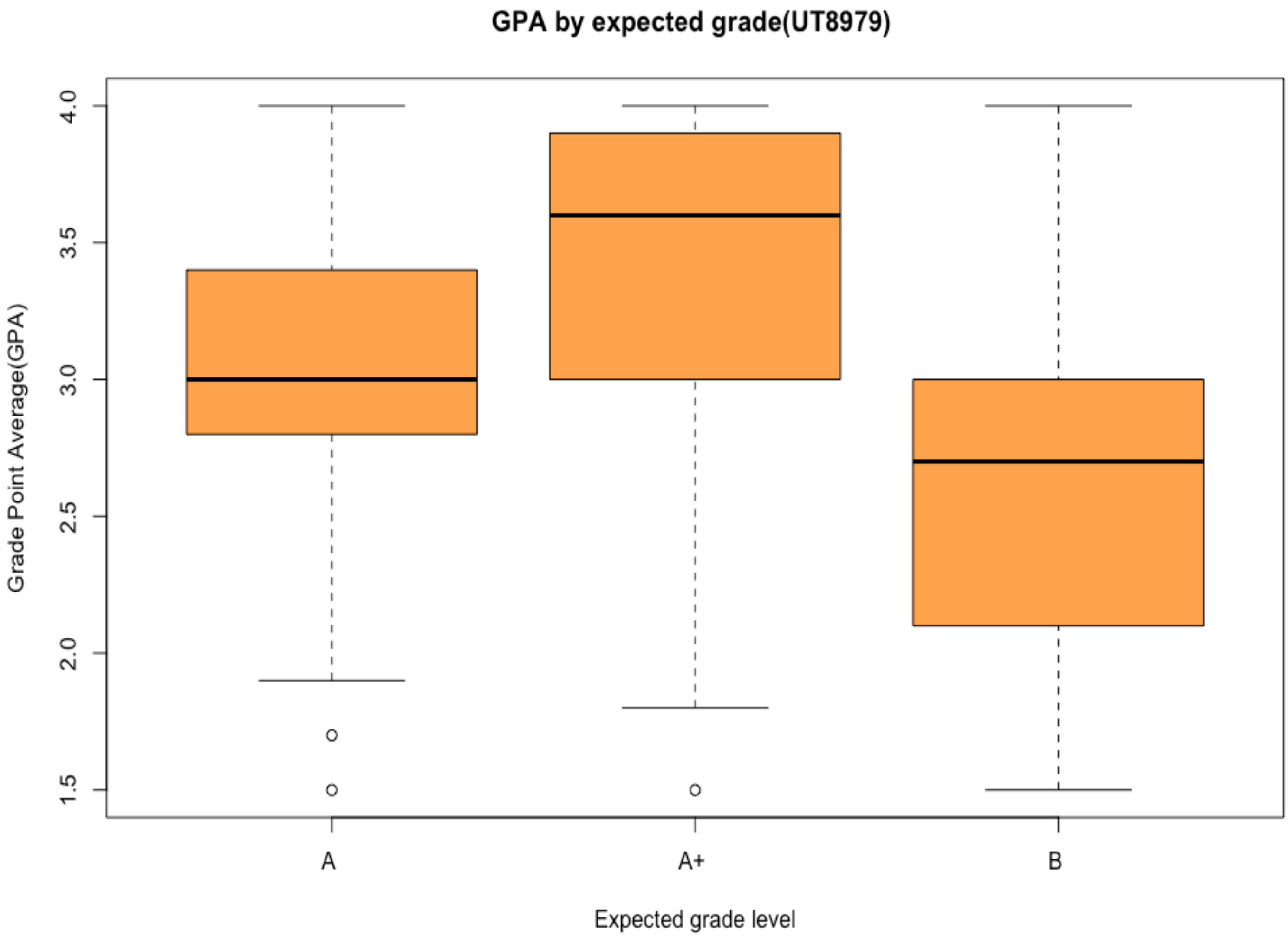
Yongwen Tan, 1002158979

1)

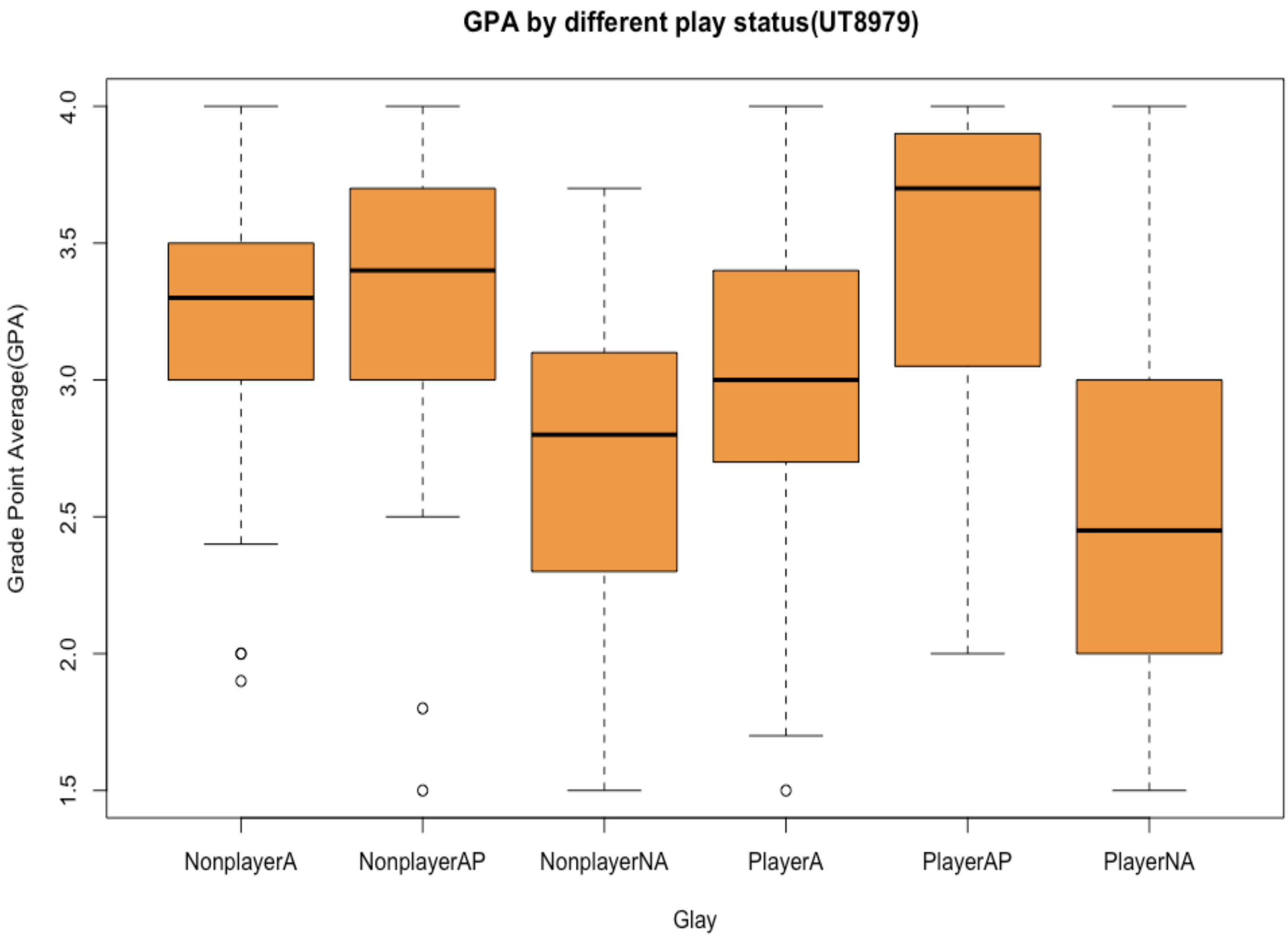
i. From the boxplots below, we can see there is a slight difference in GPA between the students who is players of video games and nonplayers. The GPA of students who are nonplayers are slightly higher than players on average.



ii. The boxplot shows that there is a big difference of GPA when the expected GPA is different because we can observe that the mean of each group is not close. We can conclude that students which with higher GPA which is above 3.0 are expecting a higher GPA in this course.



iii. We can first look at the two big groups which is players and nonplayers. No matter players or nonplayers who GPA is 3.0 or above 3.0 are expected the GPA A or A+. Students who expected A+ have the higher GPA than the students expected A and B. Thus, the trend is increasing in the boxplot which means the students with higher GPA are expecting higher grade no matter the student is player or nonplayer.



2.

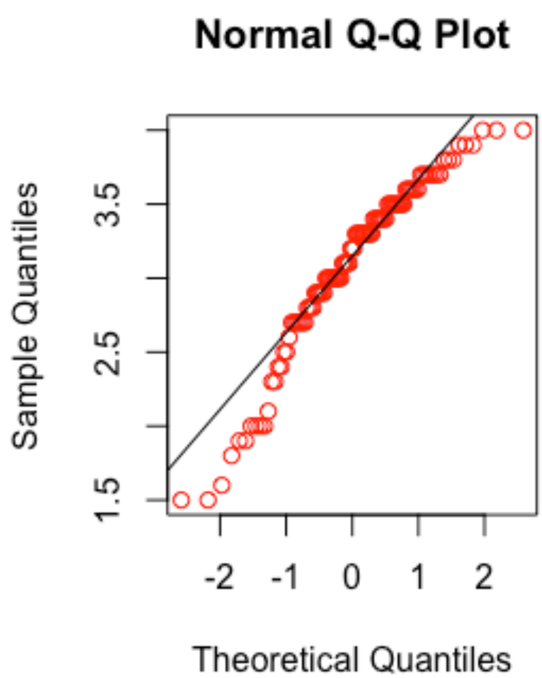
Assumption:

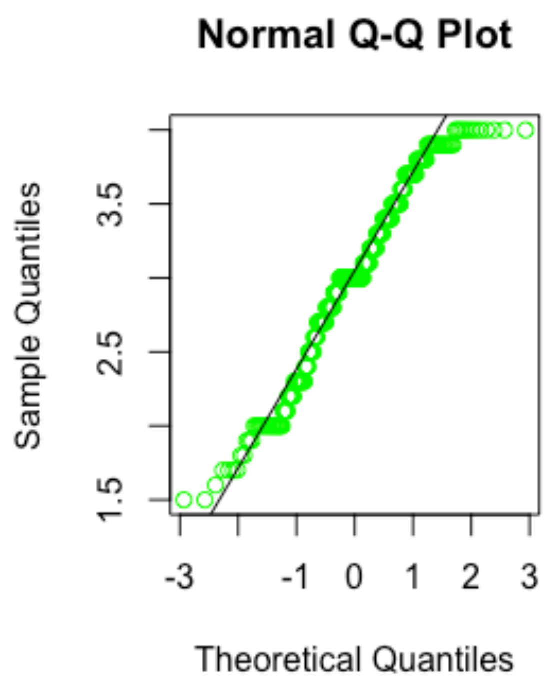
1. The data sets are independent and normally distributed.
2. The variances are equal.

Check assumption:

Use variance test for the checking the equal variance. The p-value from F test is 0.4992 which is greater than 0.05. Then, we We failed to reject Null Hypotheses. Hence, the variances are equal.

The Normal Q-Q Plots show that two samples are normal although the outliers still exist.





Pool t-test:

$$H_0: \mu_{player} = \mu_{nonplayer}$$
$$H_a: \mu_{player} \neq \mu_{nonplayer}$$

Since the p value is 0.2506 which is significantly bigger than 0.05. Thus, we failed to reject H_0. We can conclude that there is no difference in GPA between players and nonplayers.

3.

$$H_0: \mu_{GPAA+} = \mu_{GPAA} = \mu_B$$
$$H_0: \mu_{GPAA+} \neq \mu_{GPAA} \neq \mu_B$$

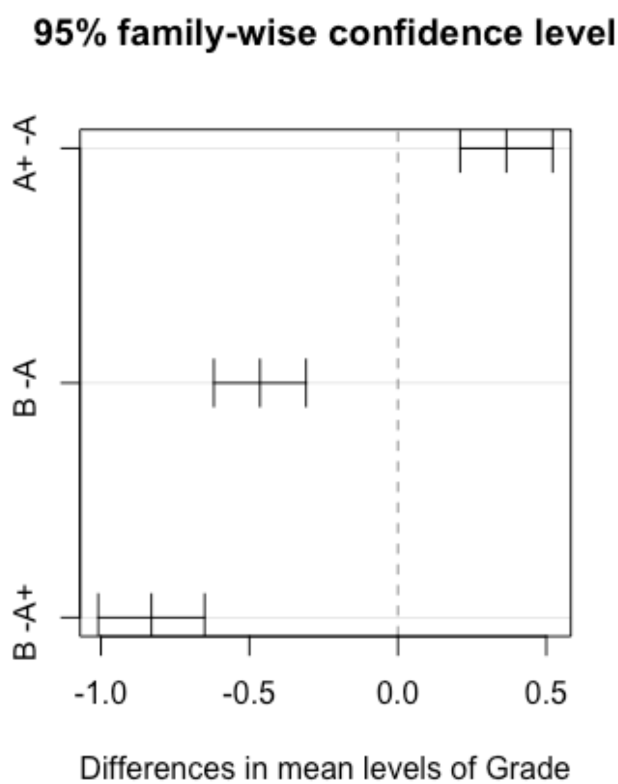
From the p-value from the anova table, which is significant and smaller than 0.05. Then we need to reject H_0. Hence, we have conclusion that the mean of at least two groups are different.

Next, we test each two groups respectively by using Tukeys approach. The results are in the table below:

Expected Grades	Confidence interval	P-values
A+ – A	(0.2097342, 0.5212961)	2e-07
B – A	(-0.6195941, -0.3090588)	0
B – A+	(-1.0089748, -0.6507084)	0

All the p-values here are greater than 0.05 which are not significant. Then we have strong evidences against the null hypothesis. In conclusion, the mean of every different level of expected grades are not equal to each other.

Also, from the plot below every confidence interval does not contain 0 which means the mean of each level of expected grades is different.



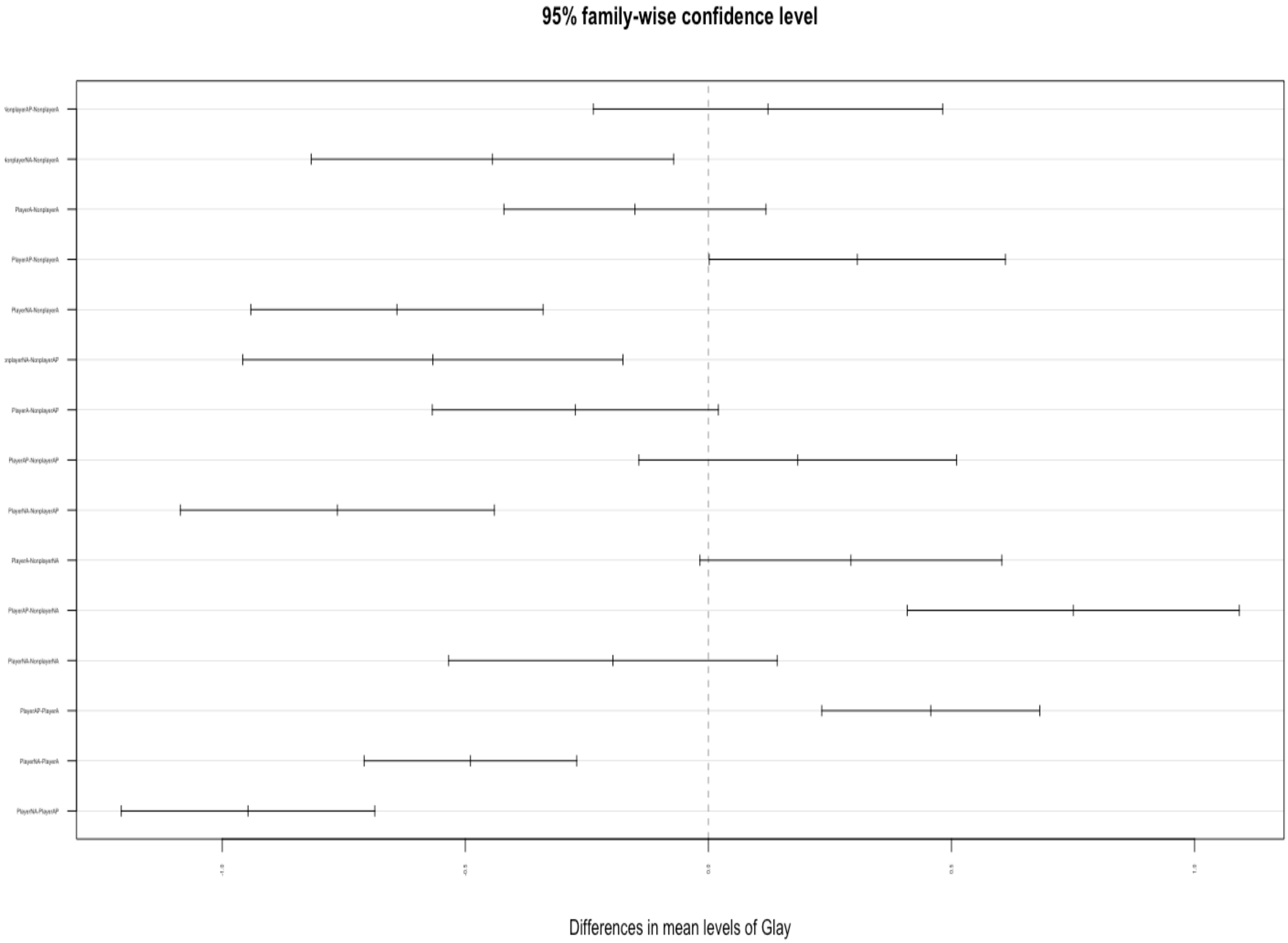
4. To concern about the difference between players and nonplayers who are expected the same level grade.

$$H_0: \mu_{GPAplayA} = \mu_{GPANnonplayerA}$$
$$H_0: \mu_{GPAplayA} \neq \mu_{GPAnonplayerA}$$
$$H_0: \mu_{GPAplayA+} = \mu_{GPANnonplayerA+}$$
$$H_0: \mu_{GPAplayA+} \neq \mu_{GPAnonplayerA+}$$
$$H_0: \mu_{GPAplayB} = \mu_{GPANnonplayerB}$$
$$H_0: \mu_{GPAplayB} \neq \mu_{GPAnonplayerB}$$

By using Tukeys approach. The result is listed below:

Glay	Confidence interval	p-value
PlayerA–NonplayerA	(-0.1510952 -0.420523778)	0.5950421
PlayerAP–NonplayerAP	(0.1837178 -0.142989193)	0.5921294
PlayerNA–NonplayerNA	(-0.1963602 -0.534229046)	0.5562541

All the p-values shown from the table above are greater than 0.05 which are non-significant. Then we fail to reject null hypothesis. We conclude that the means are the same.



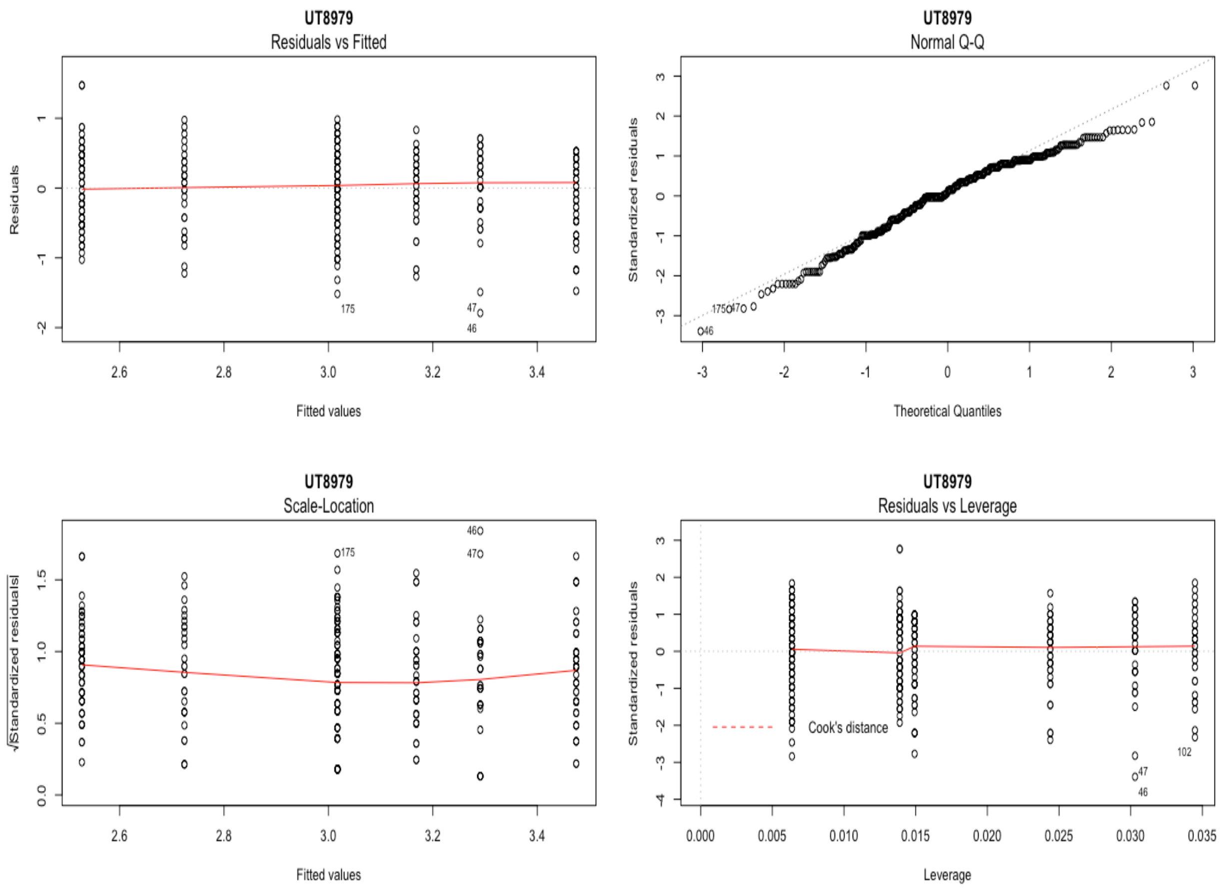
From the plot above, we can find out the confidence intervals for PlayerA-NonplayerA, PlayerAP–NonplayerAP, PlayerNA–NonplayerNA contain 0 which also means that the mean of each compared group are the same. In conclusion, for the same level of expected grades students, there is no difference for playing games or not playing games.

5.Base on the plots below,

i. Residuals versus Fitted values: I don’t see any distinctive pattern. Thus, the variance is constant which fits our assumption.

ii. Normal QQ plot: the plots show the plot are following a straight line which means the residuals are normally distributed.

iii. Square root of Standardized absolute residuals versus Fitted values: This plot also proves the equal variance assumption.



In conclusion, these three plots all show the model assumptions are satisfied. So, we can trust the conclusions. Also, the data are randomly collected which the assumption of independent holds.

We need to consider the number of students in different groups because the sample size is not big enough to stand for the common situation.

But for the thumb for variance, 0.3290394/ 0.2427195=1.356 which is smaller than 2. Then the variances are equal.

6.

a). $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \epsilon_i$

b). There are six levels of Glays in the model of question 4. Then there are 5 predictor variables. But in two-way model, there are two levels for playing video games which is play and nonplay. For this case, we need 1 predictor variable for this group. For the expected grades groups, there are three levels which is A+, A, B. Then we need two predictor variables. Also, the interaction of two groups need to be considered. So, totally there will be 5 predictor variables which is the same as the one-way model.

c). For the same level but different play status the F-test will be non-statically significant, we can get that from the results from question 4 which is for the same level of expected grades students, there is no difference for playing games or not playing games.

7.

When Play as a quantitative, it can show the exact numbers of hours that playing video games. It is easy to obverse the impact of the students' GPA.

When Play as a factor, we cannot find out the impact of every hour on the GPA. We will not be able to find out the detail.

Two-way ANOVA (with Play):

$$Y = \beta_0 + \beta_1 X_{A,i} + \beta_2 X_{B,i} + \beta_3 X_{play,i} + e_i \ , i = 1, \dots ,400$$

ANCOVA (with Play hours):

$$\mathrm{Y} = \alpha_0 + \alpha_1 X_{hours,i} + \alpha_2 X_{grades,i} + e_i \ , i = 1, \dots ,400$$

8.

If the students attend to the lectures (yes, no)

The sex of the students (male, female)

Appendix

```
> #1

> studentdata=read.csv("~/desktop/STA303/A2/data2.csv")

> attach(studentdata)

The following objects are masked from studentdata (pos = 4):

    GPA, Grade, Play

The following objects are masked from studentdata (pos = 5):

    GPA, Grade, Play

The following objects are masked from studentdata (pos = 6):

    GPA, Grade, Play

The following objects are masked from studentdata (pos = 7):

    GPA, Grade, Play

The following objects are masked from studentdata (pos = 9):
```

```
        GPA, Grade, Play

The following objects are masked from studentdata (pos = 10):

        GPA, Grade, Play

The following objects are masked from studentdata (pos = 11):

        GPA, Grade, Play

The following objects are masked from studentdata (pos = 12):

        GPA, Grade, Play

> Player=array(0,399)
> Glay<-NULL
> for (i in 1:399)
+ { if (Play[i]>0)
+ {Player[i]=1}
+       else {Player[i]=0}
+ }
> 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,
+             main="GPA for players and nonplayers(UT8979) ",
+             xlab="Students play vedio games status",
+             names = c("Nonplayers","Players"),
+             ylab="Grade Point Average(GPA)", col="tan")
> boxplot(GPA~Grade,main="GPA by expected grade(UT8979) ",
+             xlab="Expected grade level",
```

```
+                 ylab="Grade Point Average(GPA)", col="tan1")
> boxplot(GPA~Glay,main="GPA by different play status(UT8979) ",
+                 xlab="Glay",
+                 ylab="Grade Point Average(GPA)", col="tan2")
>
>
> #2
> par(mfrow=c(1,2))
> qqnorm(GPA[Player=="0"],col="red")
> qqline(GPA[Player=="0"])
> qqnorm(GPA[Player=="1"],col="green")
> qqline(GPA[Player=="1"])
> var.test(GPA~Player)


        F test to compare two variances


data:    GPA by Player
F = 0.89109, num df = 102, denom df = 295, p-value = 0.4992
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  0.6553867 1.2424011
sample estimates:
```

```
ratio of variances

           0.8910886



> t.test(GPA~Player, var.equal=TRUE)



        Two Sample t-test



data:   GPA by Player

t = 1.1506, df = 397, p-value = 0.2506

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

 -0.05728805  0.21895821

sample estimates:

mean in group 0 mean in group 1

       3.082524        3.001689



>

>

>

> #3

> model1=lm(GPA~Grade)

> anova(model1)
```

```
Analysis of Variance Table


Response: GPA
           Df  Sum Sq Mean Sq F value    Pr(>F)
Grade       2  34.867 17.4337   59.84 < 2.2e-16 ***
Residuals 396 115.370  0.2913
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(aov(GPA~Grade))
  Tukey multiple comparisons of means
    95% family-wise confidence level


Fit: aov(formula = GPA ~ Grade)


$Grade
            diff        lwr        upr p adj
A+ -A  0.3655152  0.2097342  0.5212961 2e-07
B -A  -0.4643264 -0.6195941 -0.3090588 0e+00
B -A+ -0.8298416 -1.0089748 -0.6507084 0e+00


> plot(TukeyHSD(aov(GPA~Grade)))
>
```

```
>

> #4

> model2=lm(GPA~Glay)

> anova(model2)

Analysis of Variance Table


Response: GPA

           Df  Sum Sq Mean Sq F value    Pr(>F)

Glay        5  37.153  7.4306  25.823 < 2.2e-16 ***

Residuals 393 113.084  0.2877

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> TukeyHSD(aov(GPA~Glay))

  Tukey multiple comparisons of means

    95% family-wise confidence level


Fit: aov(formula = GPA ~ Glay)


$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


> plot(TukeyHSD(aov(GPA~Glay)),las=2,cex.axis=0.33)
> tapply(GPA,Glay,var)
 NonplayerA NonplayerAP NonplayerNA     PlayerA    PlayerAP    PlayerNA
  0.2427195   0.3402273   0.3290394   0.2684844   0.2910131   0.3124570
>
>
> #5
> par(mfrow=c(2,2))
```

```
> plot(model2,main = "UT8979")

>
```