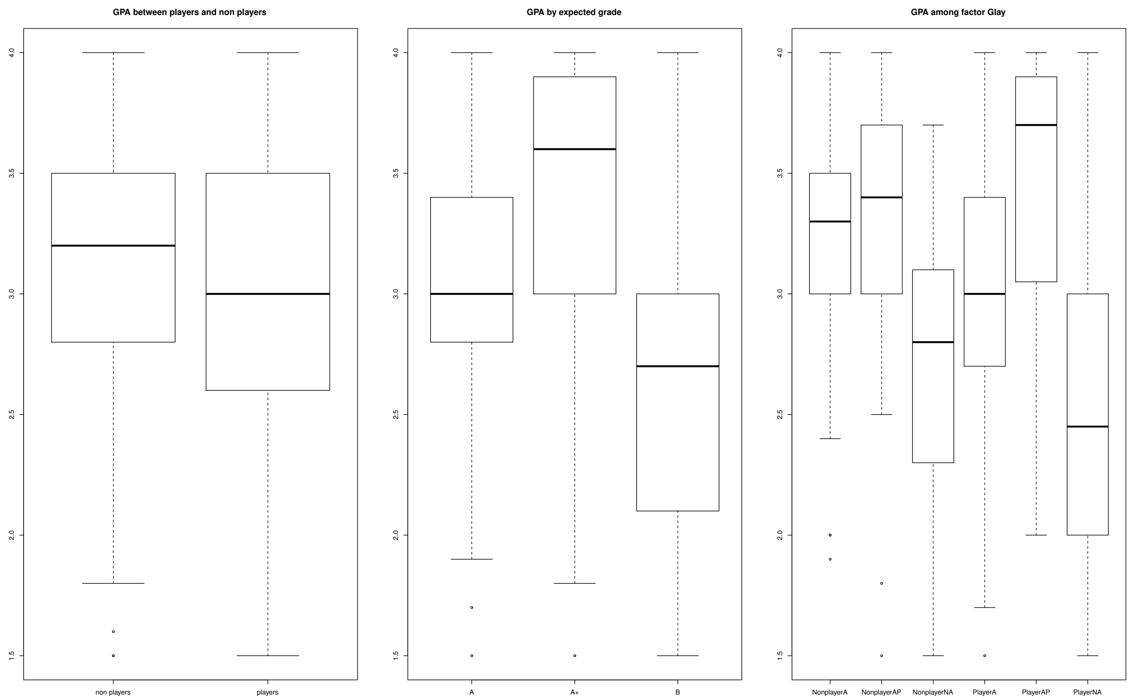
University of Toronto

GPA and grade expectation & game playing analysis

YiFei Gu

1. Construct the three side by side plot as shown below:



   The three plot appear to be different due to the size of levels of factors.
   In the first plot, `Player` is a factor with 2 levels: 0 - non player and 1 - player;
   In the second plot, `Grade` is a factor with 3 levels: A, A+, B;
   In the third plot, `Glay` is a factor with 6 levels: NonplayerA, NonplayerAP, NonplayerNA, PlayerA, PlayerAP, PlayerNA.

2. With the **two sample t-test** procedure, the hypothesis is:

$$H_0 : \mu_{\texttt{player\_gpa}} - \mu_{\texttt{non\_player\_gpa}} = 0, H_A : \mu_{\texttt{player\_gpa}} - \mu_{\texttt{non\_player\_gpa}} \neq 0$$

   `t.test` gives an output with `p-value = 0.2383`, as shown in the **Appendix**. The p-value does not give evidence to reject the null hypothesis, which implies that there is not a significant difference in the mean of GPA between the player and non player of video and/or computer games.

3. From **one-way analysis of variance** with

$$H_0 : \mu_{\texttt{A\_gpa}} = \mu_{\texttt{A+\_gpa}} = \mu_{\texttt{B\_gpa}},\ H_A : \exists\ i \neq j \ \text{ s.t } \ \mu_i \neq \mu_j$$

the summary of the one way anova function: `summary(aov())` gives an output with `p-value <2e-16` as shown in the **Appendix**, which provides significant evidence to reject the null hypothesis of equal means.
In order to find which levels grades differ, perform **pair-wise t test** with Bonferroni's correction method, with `pairwise.t.test(gpa, grade, p.adj = "bonf")`, we get a table of statistical significance between each pairs:

	A	A+
A+	1.8e-07	-
B	2.6e-11	<2e-16

Which suggests that the difference of means of gpa between group with expected grade of A and A+, A and B, A+ and B are all significantly difference from each other, with `p-value` of `1.8e-07, 2.6e-11, 2e-16` respectively.

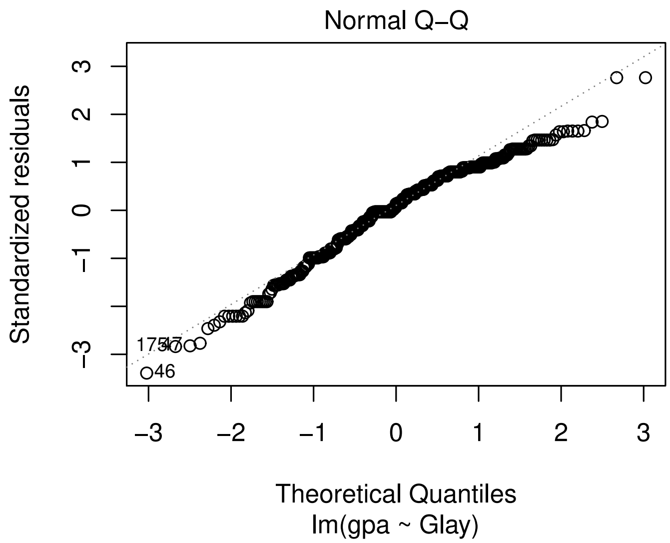
4. Let μ_i denote the mean of gpa of ith category in the six category of of students classified by the combination of their player status and expected grade. Then the null hypothesis is:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6,\ H_A : \exists\ i \neq j \ \text{ s.t } \ \mu_i \neq \mu_j$$

As shown in **Appendix** section, the `p-value` from the output is less than `2e-16`, which provides strong evidence to reject the null hypothesis of equal means.
With the **Tukey's Honest Significance Test**, we can also compare the statistical significance between each pair in the group. The following table provides a detailed look for pairs with significant difference, which can also be find in the **Appendix**:

	p adj
NonplayerNA-NonplayerA	0.0092179
PlayerAP-NonplayerA	0.0477882
PlayerNA-NonplayerA	0.0000000
NonplayerNA-NonplayerAP	0.0005766
PlayerNA-NonplayerAP	0.0000000
PlayerAP-NonplayerNA	0.0000000
PlayerAP-PlayerA	0.0000002
PlayerNA-PlayerA	0.0000000
PlayerNA-PlayerAP	0.0000000

5. (a) In order to ensure the results from above tests are valid, several assumptions about the model need to be checked: One is **Homoscedasticity**, Second is **Normality**.
With **Bartlett's test**, one can access assumption of equal variance. From the function `bartlett.test()`, `p-value = 0.8644` is obtained as shown in **Appendix**, which provides a conclusion that there is no evidence to reject the hypothesis of equal variance.
The following qq-plot suffice to access the assumption of normality:



As shown above, the assumption of normality roughly holds, thus one can conclude the previous tests are valid.

(b) We should concern the issue, since unequal sample sizes can affect the homogeneity of variance assumption, especially for small sample sizes. Since homogeneity is already checked, then there is no need to concern for the results above. The larger and closer sample size for each group, the better for analysis.

6. (a) Let the Y_i denotes the GPA on the i-th row, $\mathbb{1}_{\text{play},i}$ indicator variable denoting whether play games or not. $\mathbb{1}_{A,i}\&\mathbb{1}_{A+,i}\&\mathbb{1}_{B,i}$ denoting the expected grade, then:

$$Y_i = \beta_0+\beta_1\mathbb{1}_{\text{play},i}+\beta_2\mathbb{1}_{A,i}+\beta_3\mathbb{1}_{A+,i}+\beta_4\mathbb{1}_{B,i}+\beta_5\mathbb{1}_{\text{play},i}\times\mathbb{1}_{A,i}+\beta_6\mathbb{1}_{\text{play},i}\times\mathbb{1}_{A+,i}+\beta_7\mathbb{1}_{\text{play},i}\times\mathbb{1}_{B,i}+\epsilon_i$$

(b) The total number of predictors increases since two-way anova also includes the independent factor as predictors along with interaction term.

(c) The null hypothesis for F-test is

$$H_0 : \beta_1 = \beta_2 = ... = \beta_{\text{df model}}$$

which implies non of the factors contribute to the response variable in the general linear model.

Since results from question 4 implies that there is significant difference of mean between a subset of groups, this implies that some predictor variable in question 4 does affect GPA in a statistical significant way, which implies that the for sure a subset of β's will differ from each other, thus the F-test will be statistically significant to reject the null hypothesis.

7. When using **Play** as a quantitative explanatory variable, one wants to find if the change of play time affect GPA; That is, if y is the GPA and x as a quantitative explanatory variable that denotes play time, one can check the statistical significance of β's in a linear model:

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + ... + \epsilon_i$$

When as a factor in an additive model, play is treated as categorical factor that one only whether play games or not have an effect on GPA is investigated.

That is, **Play** is used as an indicator variable $\mathbb{1}_{\text{play}}$, with a hypothesis that the coefficient of this indicator variable $\beta_i = 0$, and thus will only have a treatment effect of the value β_i, which can possiblly shift response variable vertically.

8. The following indicator variable can potentially influence GPA:
   Let $\mathbb{1}_g$ denote whether a student wear a glasses or not, with two levels: 0 - do not wear, 1 - wear.
   $\mathbb{1}_f$ denote if a student is in a relationship, also with two levels: 0 - no, and 1 - yes.

Appendix:

1.
```
> par(mfrow=c(1,3))
> boxplot(gpa~Player, main = 'GPA between players and non players',
names=c("non players","players"))
> boxplot(gpa~Grade, main = 'GPA by expected grade')
> boxplot(gpa~Glay, main = 'GPA among factor Glay')
```

2.
```
> np = data[which(Play == 0),]
> p = data[which(Play > 0),]
> t.test(p$GPA, np$GPA)

        Welch Two Sample t-test

data:  p$GPA and np$GPA
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:
 -0.21561458  0.05394441
sample estimates:
mean of x mean of y
 3.001689  3.082524
```

3.
```
> aov(gpa~Grade)
Call:
   aov(formula = gpa ~ Grade)

Terms:
                   Grade Residuals
Sum of Squares  34.86739 115.36960
Deg. of Freedom        2       396

Residual standard error: 0.5397568
Estimated effects may be unbalanced
> result = aov(gpa~Grade)
> summary(result)
             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, 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
```

4.
```
> rGlay = aov(gpa~Glay)
> summary(rGlay)
             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
> TukeyHSD(rGlay, conf.level = 0.95)
  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
```

5.
```
> bartlett.test(gpa~Glay)

        Bartlett test of homogeneity of variances

data:  gpa by Glay
Bartlett's K-squared = 1.8885, df = 5, p-value = 0.8644
# for the qq-plot:
> plot(lm(gpa~Glay), which = 2)
```