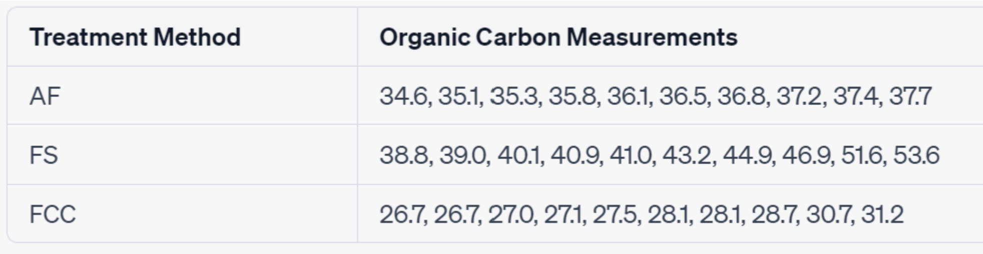
Homework 3 ANOVA

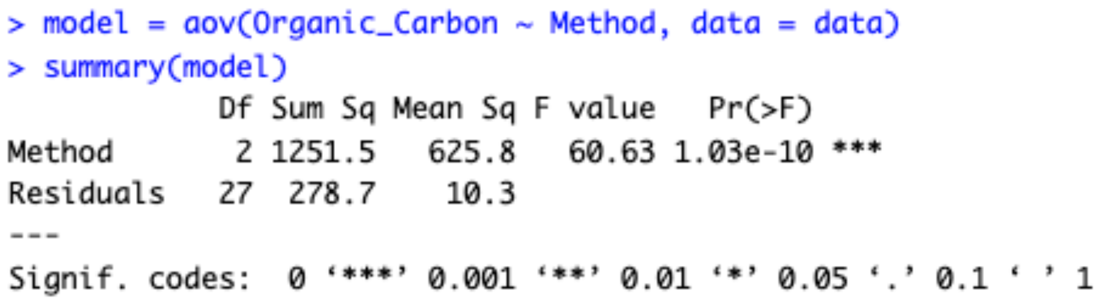
1. Data: Waste Water
   Let group 1 = “AF”, group 2 = “FS”, group 3: “FCC”
   Scientists concerned with treatment of tar sand wastewater studied three treatment methods for the removal of organic carbon. (Based on W. R. Pirie, Statistical Planning and Analysis for Treatments of Tar Sand Wastewater, Technical Information Center, Office of Scientific and Technological Information, United States Department of Energy.) The three treatment methods used were air flotation (AF), foam separation (FS), and ferric-chloride coagulation (FCC). The organic carbon material measurements for the three treatments yielded the following data:



Treatment Method	Organic Carbon Measurements
AF	34.6, 35.1, 35.3, 35.8, 36.1, 36.5, 36.8, 37.2, 37.4, 37.7
FS	38.8, 39.0, 40.1, 40.9, 41.0, 43.2, 44.9, 46.9, 51.6, 53.6
FCC	26.7, 26.7, 27.0, 27.1, 27.5, 28.1, 28.1, 28.7, 30.7, 31.2

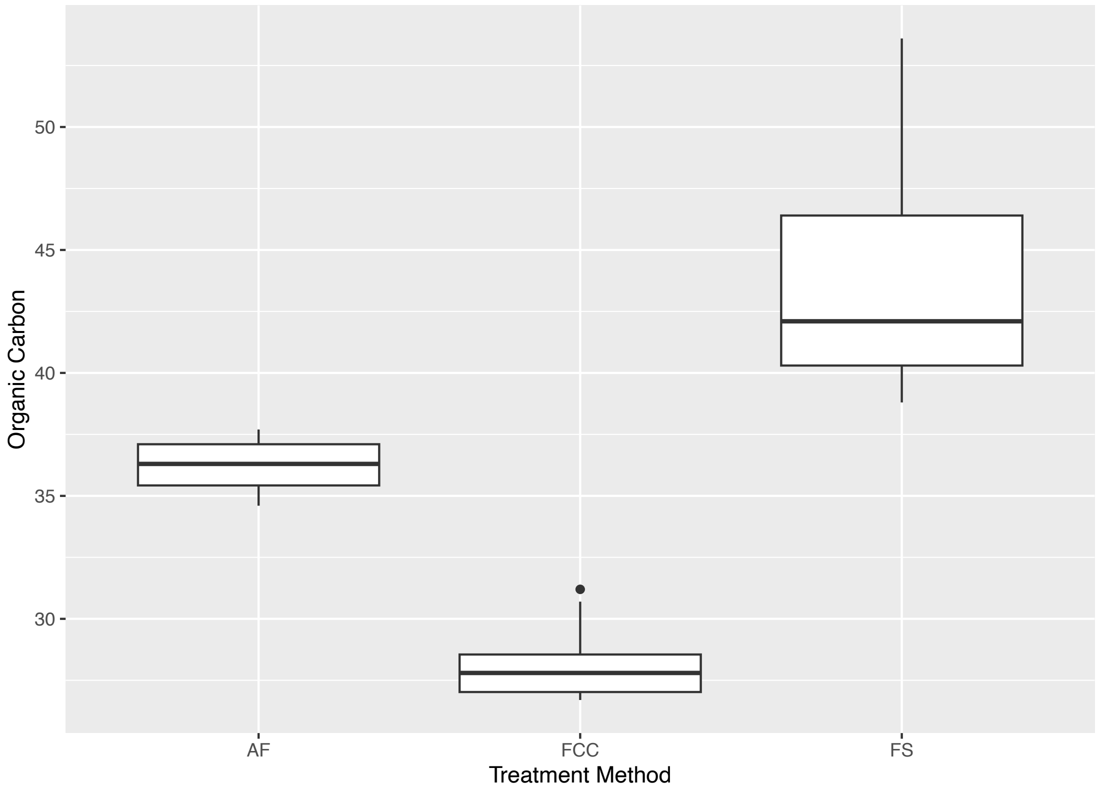
The data is provided in the file “wastewater.csv”.

a. Test $H_0 : \mu_1 = \mu_2 = \mu_3$ versus $H_a :$ not H_0 at 5% level of significance. State your conclusion.
Hint: One-Way ANOVA
Ans : 可以看到 p-value 非常小(< 0.05)表示三種方法之間存在顯著差異，因此結論為拒絕虛無假設 H_0。



```
> model = aov(Organic_Carbon ~ Method, data = data)
> summary(model)
            Df Sum Sq Mean Sq F value   Pr(>F)    
Method       2 1251.5   625.8   60.63 1.03e-10 ***
Residuals   27  278.7    10.3                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

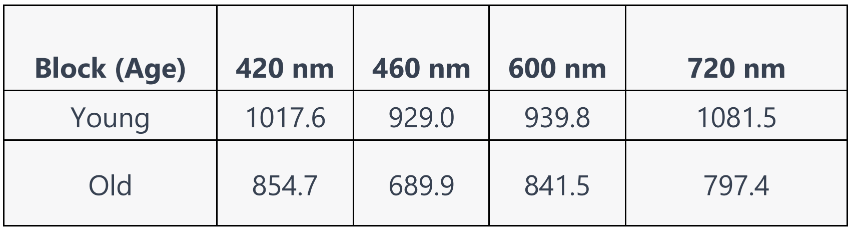
b. Plot side-by-side boxplots for the three groups and comment on the results. Which method is the best?
Ans : AF 和 FCC 的資料分散較集中而 FS 資料較分散，但 FS 資料整體分數比其他兩種方法高，且可以看到 FCC 中有一個 outlier，可以看到 FS 方法明顯優於其他兩個方法。



2. Data = “fern.csv”

研究光的波長對蕨類生長的影響

Astudy is conducted of the effect of light on the growth of ferns. Since plants grow at various rates at different ages. this variable is controlled by blocking. Four young plants (plants grown in the dark for 4 days) and four older plants (plants grown in the dark for 12 days) are utilized in the study. thus producing two blocks each of size 4. Four different light treatments are investigated. Each treatment is randomly assigned to one plant in each block. The treatments consist of exposing-each plant to a single dose of light, returning it to the dark, and measuring the cross-sectional area of the fern tip 24 hours after the light is administered. These data resulted (cross-sectional area is given in square micrometers):

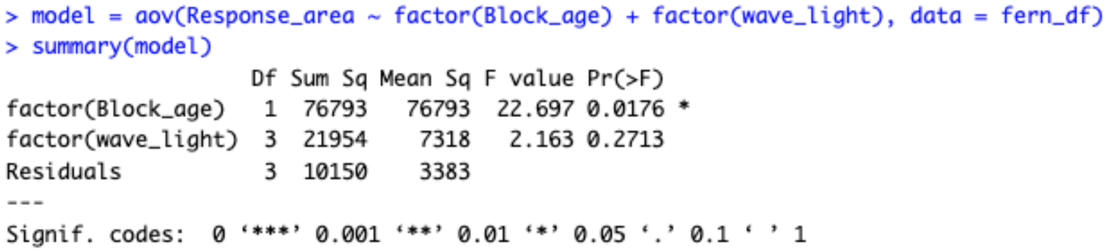


Block (Age)	**420 nm**	**460 nm**	**600 nm**	**720 nm**
Young	1017.6	929.0	939.8	1081.5
Old	854.7	689.9	841.5	797.4

a. What is the blocking variable? Please test whether the blocking effect exists or not at 5% level of significance. State your conclusion.

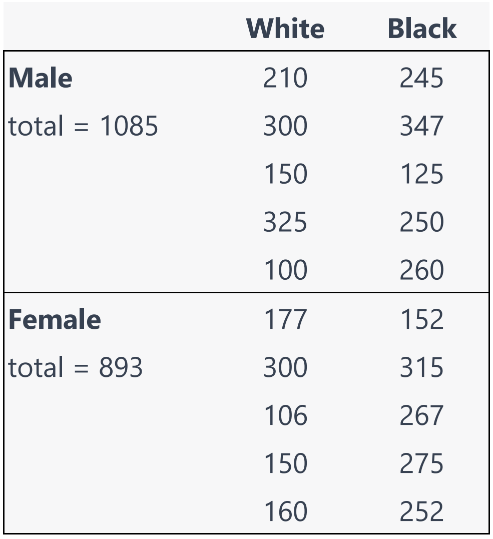
Ans : blocking variable 是一個影響結果的變數，但是不想將其視為研究的因素，希望”組間的差異”完全來自 treatment effect；由下圖可以看出 Block_age 的 p-value 小於 0.05，因此拒絕虛無假設，表示存在 blocking effect。

b. Please test whether the treatment effect (i.e. wavelength of light) exists or not at 5% level of significance. State your conclusion.
Ans： 由下圖可以看出 wave_light 的 p-value 大於 0.05，因此無法拒絕虛無假設，表示不存在 treatment effect。



```
> model = aov(Response_area ~ factor(Block_age) + factor(wave_light), data = fern_df)
> summary(model)
                   Df Sum Sq Mean Sq F value Pr(>F)  
factor(Block_age)   1  76793   76793  22.697 0.0176 *
factor(wave_light)  3  21954    7318   2.163 0.2713  
Residuals           3  10150    3383                 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

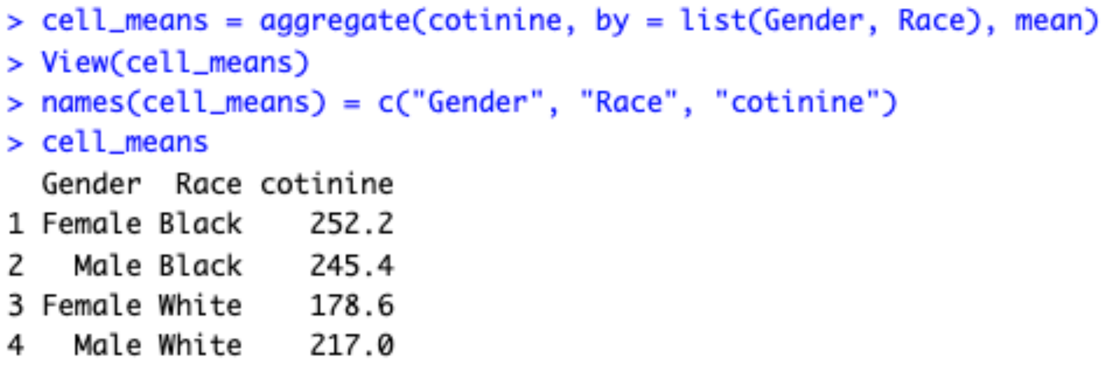
3. Data: “Cotinine.csv”
Cotinine is a major metabolite of nicotine. It is currently considered to be the best indicator of tobacco smoke exposure. A study is conducted to detect possible racial differences in cotinine level in young adults. These data are obtained on the cotinine level in milligrams per milliliter:



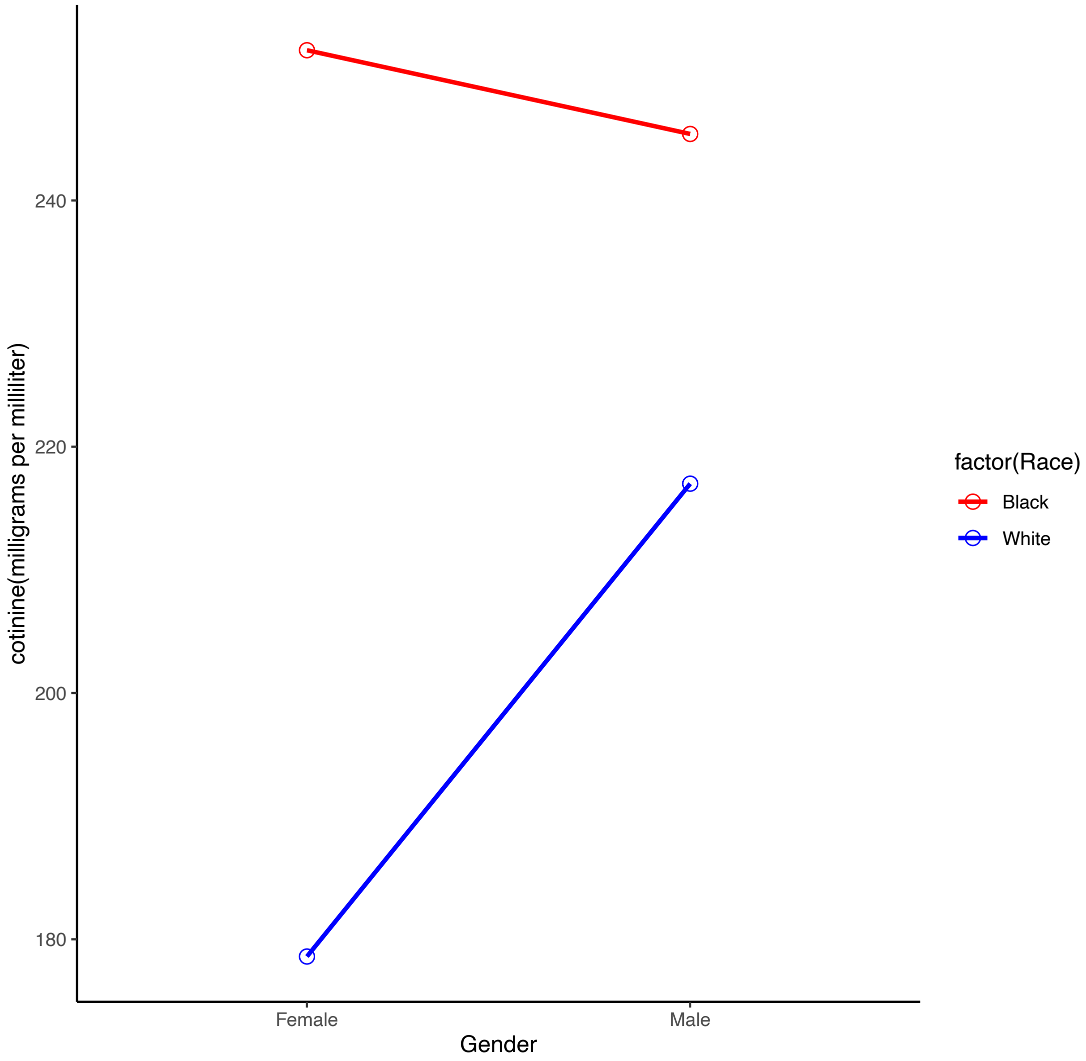
	White	Black
Male	210	245
total = 1085	300	347
	150	125
	325	250
	100	260
Female	177	152
total = 893	300	315
	106	267
	150	275
	160	252

a. Plot the means for the 4 treatment combinations. Comment on whether interaction effect exists.
Ans： 由下圖可以看出線段並不平行，表示存在 interaction effect，但

是”mean”的圖為呈現 variance 的資訊故為非正式檢定，需要再做進一步正式檢定。

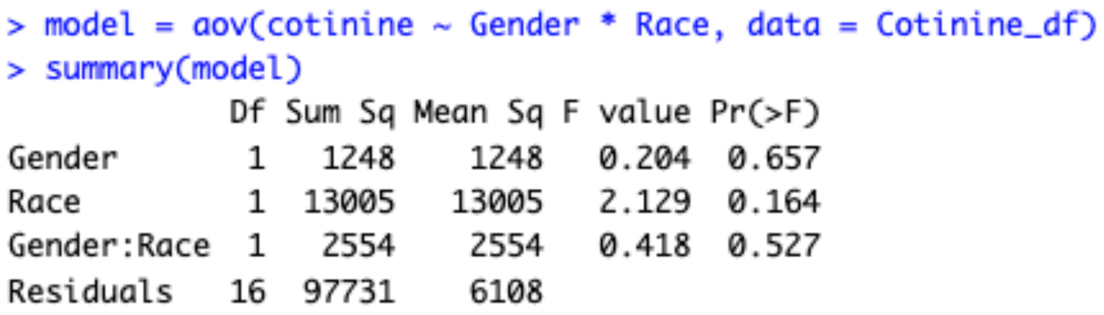


```
> cell_means = aggregate(cotinine, by = list(Gender, Race), mean)
> View(cell_means)
> names(cell_means) = c("Gender", "Race", "cotinine")
> cell_means
  Gender  Race cotinine
1 Female Black    252.2
2   Male Black    245.4
3 Female White    178.6
4   Male White    217.0
```



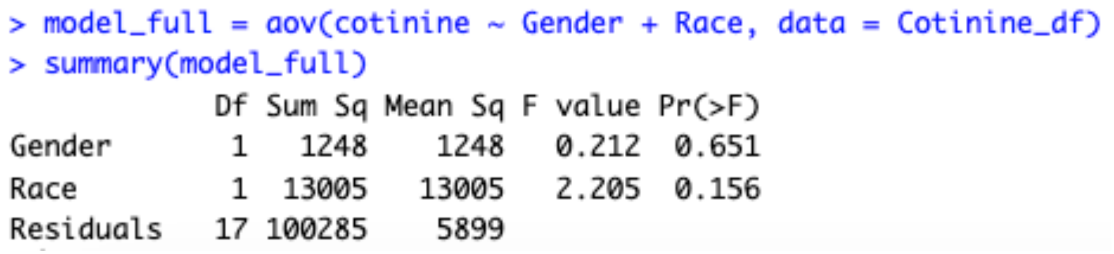
b. Perform two-way ANOVA and test whether interaction effect exists or not. Level of significance = 5%.

Ans : 由下圖中 Gender : Race 交互作用項的 p-value 大於 0.05，因此無法拒絕虛無假設，表示不存在 interaction effect。



```
> model = aov(cotinine ~ Gender * Race, data = Cotinine_df)
> summary(model)
            Df Sum Sq Mean Sq F value Pr(>F)
Gender       1   1248    1248   0.204  0.657
Race         1  13005   13005   2.129  0.164
Gender:Race  1   2554    2554   0.418  0.527
Residuals   16  97731    6108
```

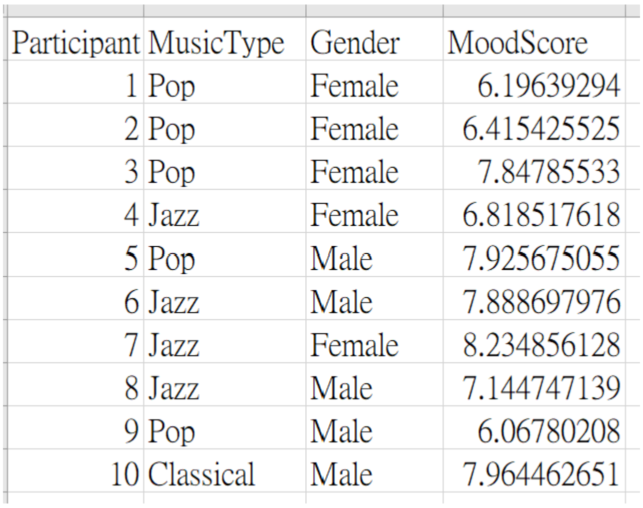
c. Test the two main effects at 5% level of significance.
Ans : 由下圖結果可以看到 Gender 和 Race 的 p-value 都大於 0.05，因此無法拒絕虛無假設，表示兩個對 cotinine 都沒有顯著影響。



```
> model_full = aov(cotinine ~ Gender + Race, data = Cotinine_df)
> summary(model_full)
            Df Sum Sq Mean Sq F value Pr(>F)
Gender       1   1248    1248   0.212  0.651
Race         1  13005   13005   2.205  0.156
Residuals   17 100285    5899
```

4. In homework 2, you have analyzed the data provided in the file “mood.csv”.

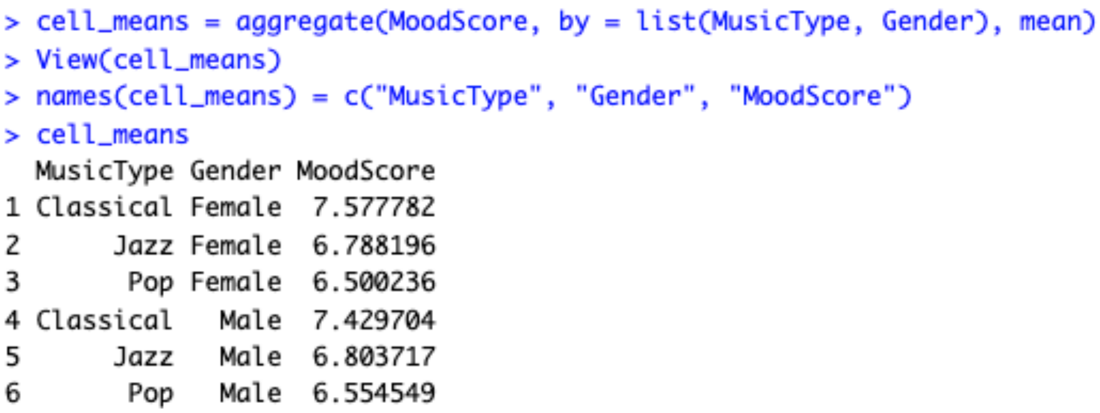
Suppose we are interested in studying the effect of different types of music on people’s moods. We collect data on 60 participants and record their mood score (out of 10) after listening to one of three types of music: classical, jazz, or pop. In the file “mood.csv” The data look like



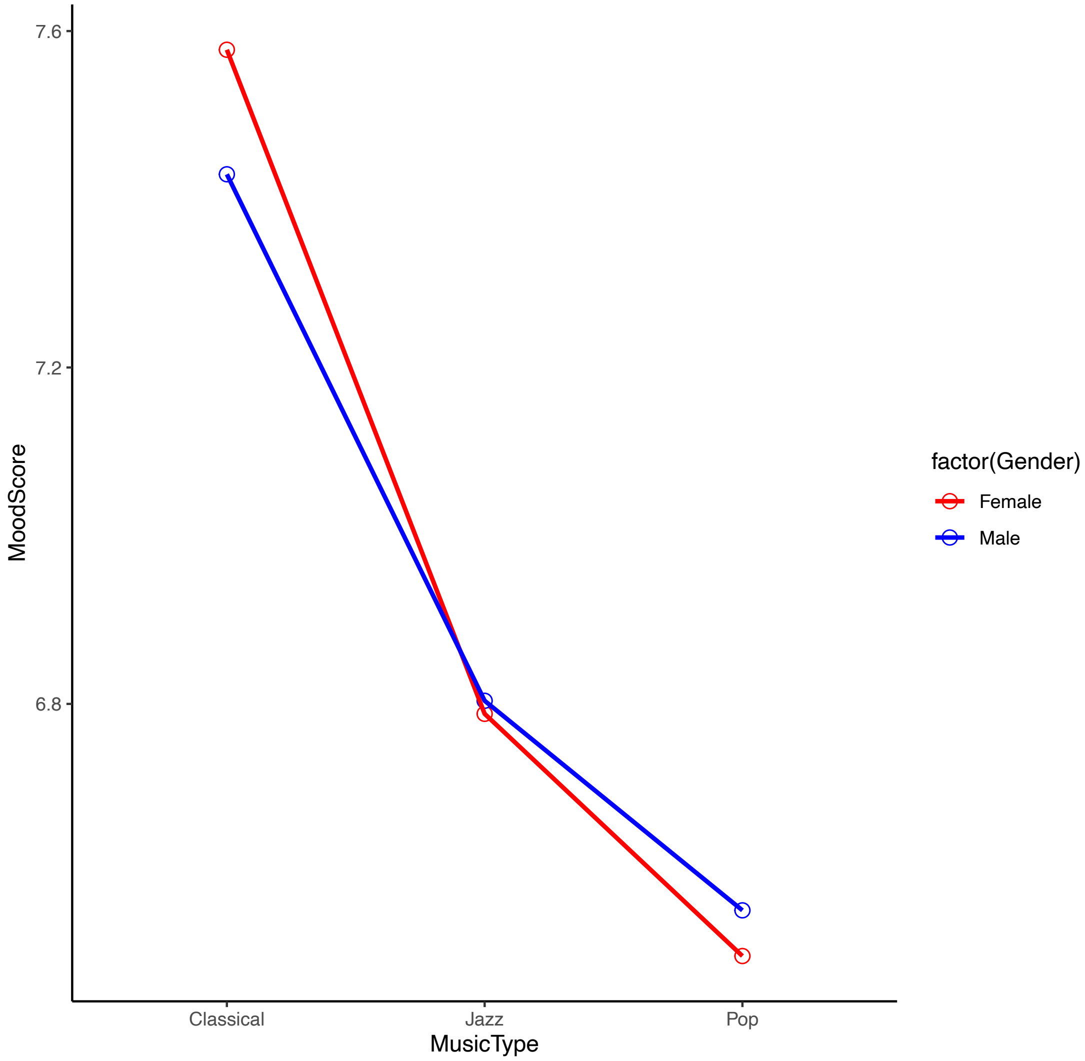
Participant	MusicType	Gender	MoodScore
1	Pop	Female	6.19639294
2	Pop	Female	6.415425525
3	Pop	Female	7.84785533
4	Jazz	Female	6.818517618
5	Pop	Male	7.925675055
6	Jazz	Male	7.888697976
7	Jazz	Female	8.234856128
8	Jazz	Male	7.144747139
9	Pop	Male	6.06780208
10	Classical	Male	7.964462651

a. Plot the means for the 6 treatment combinations. Comment on whether interaction effect exists.

Ans： 由下圖可以看出線段並不平行，表示存在 interaction effect，但是” mean” 的圖為呈現 variance 的資訊故為非正式檢定，需要再做進一步正式檢定。

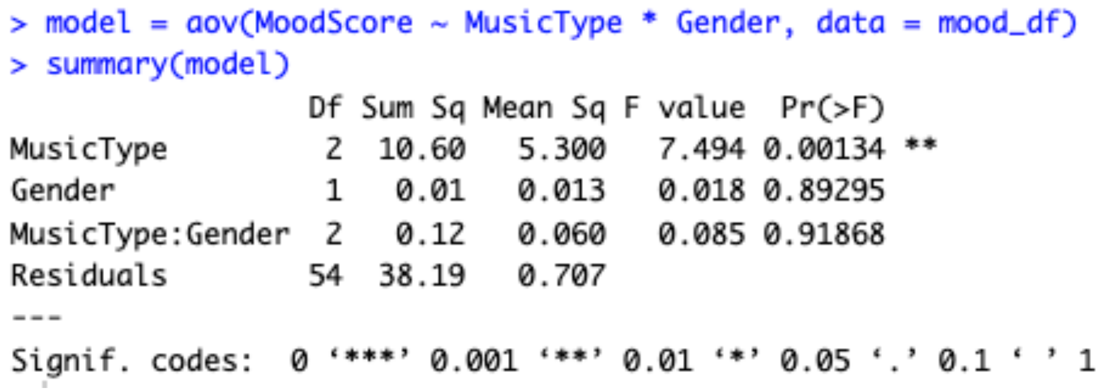


```
> cell_means = aggregate(MoodScore, by = list(MusicType, Gender), mean)
> View(cell_means)
> names(cell_means) = c("MusicType", "Gender", "MoodScore")
> cell_means
  MusicType Gender MoodScore
1 Classical Female  7.577782
2      Jazz Female  6.788196
3       Pop Female  6.500236
4 Classical   Male  7.429704
5      Jazz   Male  6.803717
6       Pop   Male  6.554549
```



b. Perform two-way ANOVA and test whether interaction effect exists or not. Level of significance = 5%. Explain the result.

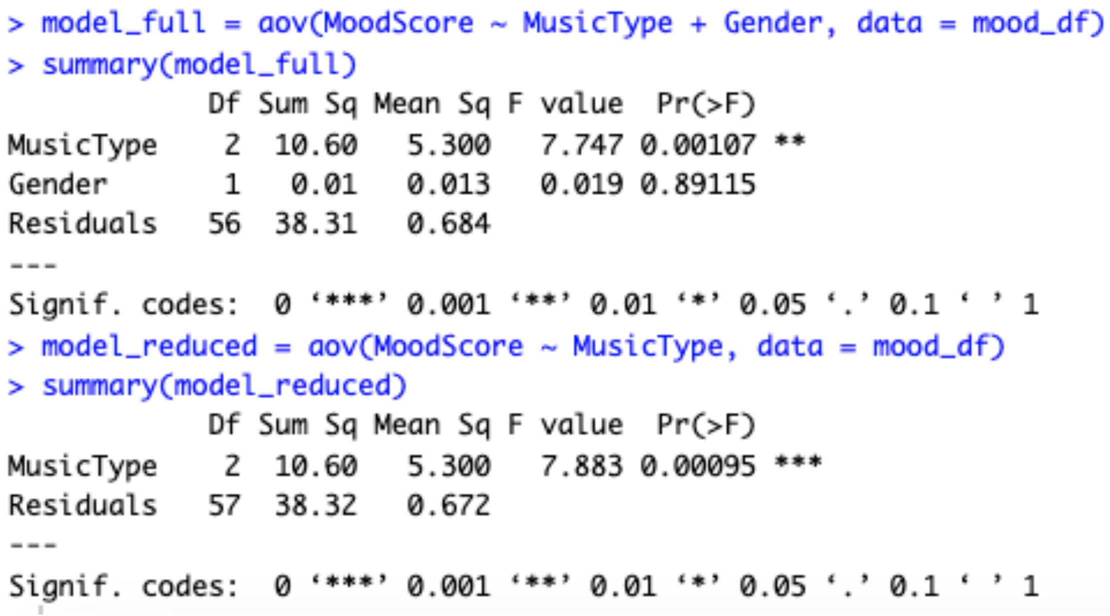
Ans : 由下圖中 MusicType : Gender 交互作用項的 p-value 大於 0.05，因此無法拒絕虛無假設，表示不存在 interaction effect，但可以看到 MusicType 的 p-value 小於 0.05，表示其對 MoodScore 有顯著影響，而 Gender 沒有顯著影響。



```
> model = aov(MoodScore ~ MusicType * Gender, data = mood_df)
> summary(model)
                 Df Sum Sq Mean Sq F value  Pr(>F)   
MusicType         2  10.60   5.300   7.494 0.00134 **
Gender            1   0.01   0.013   0.018 0.89295   
MusicType:Gender  2   0.12   0.060   0.085 0.91868   
Residuals        54  38.19   0.707                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

c. Test the two main effects at 5% level of significance. Explain the result.

Ans : 由下圖可以看到 MusicType 的 p-value 非常小(< 0.05)，表示 MusicType 對 MoodScore 有顯著影響，但是 Gender 的 p-value 大於 0.05，無法拒絕虛無假設，表示其對 MoodScore 沒有顯著影響，因此進一步對 MusicType 做 one-way anova 發現 F-value 上升和 p-value 更小了，表示 MusicType 對 MoodScore 有更顯著的影響。



```
> model_full = aov(MoodScore ~ MusicType + Gender, data = mood_df)
> summary(model_full)
            Df Sum Sq Mean Sq F value  Pr(>F)   
MusicType    2  10.60   5.300   7.747 0.00107 **
Gender       1   0.01   0.013   0.019 0.89115   
Residuals   56  38.31   0.684                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> model_reduced = aov(MoodScore ~ MusicType, data = mood_df)
> summary(model_reduced)
            Df Sum Sq Mean Sq F value  Pr(>F)    
MusicType    2  10.60   5.300   7.883 0.00095 ***
Residuals   57  38.32   0.672                    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```