Homework #2

1. Pizza was original invented in Naples, Italy in the early 19th century. It is a kind of flat bread baked by oven and is usually topped with cheese, tomato sauce, meat and vegetables. Pizza has become a common delicacy around the world.

Suppose the dataset **pizza2.txt** contains data on pizzas sold at a US pizzeria, which could provide insights into factors that influence pizza ratings. The table below shows some key information about the data.

Data	pizza2.txt	
Description	Data about pizza	
Variables descriptions	rating	Rating for the pizza
	cost	Cost per slice
	heat	Heat source used (Gas/Coal/Wood)
	brick	The use of brick oven (TRUE/FALSE)
	area	The location of pizzeria
	heat_re **這是一個錯誤的編碼**	Same as the **heat** variable but **it is just numerically coded instead of using strings.** 0 – Coal 1 – Wood 2 – Gas

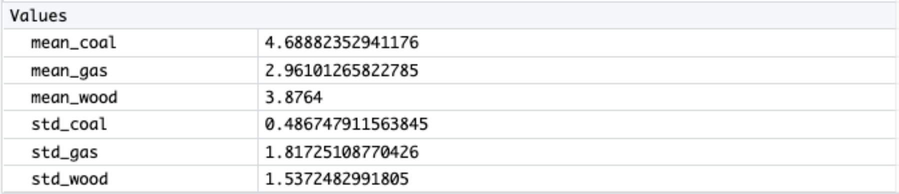
Tasks:

I. "Using coal to bake pizzas yields different ratings with those baked by using gas or wood".

We wish to verify this statement by providing some statistical evidences:

a. Compute each of the average ratings of the pizzas baked by coal, wood and gas, along with the standard deviations of the ratings. Comment the results. *[hint: you could use codes like **pizza[pizza[,"heat"]=="Coal", ratings]** **OR** **sapply()** and a self-defined function to do so]*

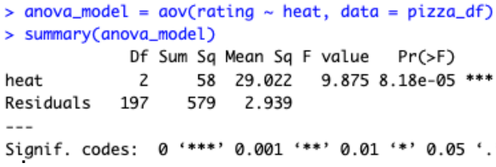
Ans：可以看到”Coal”的 rating 平均明顯高於其他兩種烘烤方法，表示用”Coal”烘烤的 pizza 平均有較高的 rating，從標準差可以看出資料的分佈情況，”Coal” 的分佈程度最集中，其次是 “Wood”，最後的是”Gas”表示其分佈程度最分散。



Values	
mean_coal	4.68882352941176
mean_gas	2.96101265822785
mean_wood	3.8764
std_coal	0.486747911563845
std_gas	1.81725108770426
std_wood	1.5372482991805

b. Perform an ANOVA test to find out if the ratings of the pizzas baked by different heat sources are equal in average. Comment the results.

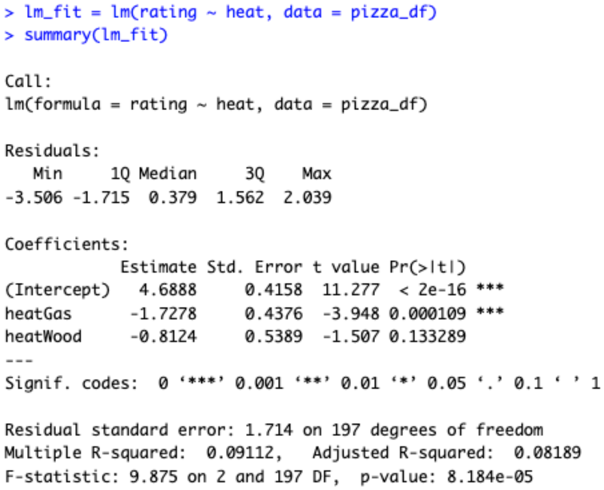
Ans : 可以看到 p-value 為 8.18e-05 小於顯著水準 0.05，則拒絕虛無假設認為不同烘烤方式之間有顯著的差異，三個星號代表非常顯著。



```
> anova_model = aov(rating ~ heat, data = pizza_df)
> summary(anova_model)
             Df Sum Sq Mean Sq F value   Pr(>F)    
heat          2     58  29.022   9.875 8.18e-05 ***
Residuals   197    579   2.939                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

c. Fit a simple linear regression by using **rating** as the response variable and **heat** as the predictor variable. Interpret the estimated regression coefficients and the corresponding p-values.

Ans : 結果表示當 heat 為”Coal”時，rating 的 Estimate 為 4.6888，p-value 為非常顯著，表示使用”Coal”的烘烤方式與 rating 之間的關係是顯著的。當 heat 變為”Gas”時，Estimate 會下降 1.7278 而其 p-value 非常小(＜0.001)，表示使用”Gas”的烘烤方式與 rating 之間的關係是顯著的。當 heat 變為”Wood”時，Estimate 則會下降 0.8124 而其 p-value 為 0.133289，表示使用”Wood”的烘烤方式與 rating 之間的關係可能沒有顯著的關聯。



```
> lm_fit = lm(rating ~ heat, data = pizza_df)
> summary(lm_fit)

Call:
lm(formula = rating ~ heat, data = pizza_df)

Residuals:
   Min     1Q Median     3Q    Max 
-3.506 -1.715  0.379  1.562  2.039 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   4.6888     0.4158  11.277  < 2e-16 ***
heatGas      -1.7278     0.4376  -3.948 0.000109 ***
heatWood     -0.8124     0.5389  -1.507 0.133289    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.714 on 197 degrees of freedom
Multiple R-squared:  0.09112,	Adjusted R-squared:  0.08189 
F-statistic: 9.875 on 2 and 197 DF,  p-value: 8.184e-05
```

d. Compare and contrast the results in (a), (b) and (c). In other words, what information are shown from both analyses, *OR* from one analysis, but not from the others?

Ans：單變量分析提供三種烘烤方式的平均和分佈情況，但是並沒有考慮多變量之間的關係，ANOVA 則是提供不同烘烤方式之間的統計差異，但是並沒有考慮其他可能的因素，迴歸分析則是提供不同烘烤方式和 rating 之間的關係，但是並沒有提供關於多變數之間的統計差異。單變量並不能得知 p-value 但能從 ANOVA 和迴歸分析得知，ANOVA 不能得知各個變數的 p-value 但迴歸分析可以。

(a 小題是單變量分析, b 小題是 ANOVA, c 小題是迴歸. 所以 d 小題要你比較三種方法得出來的結果)

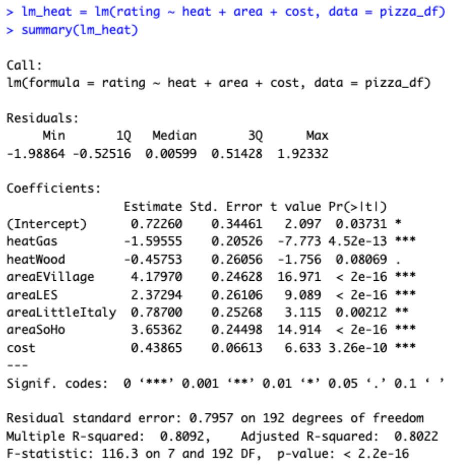
II. Fit two multiple linear regression by using **rating** as the response variable, and

e. **heat**, **area** and **cost** as the predictor variables.

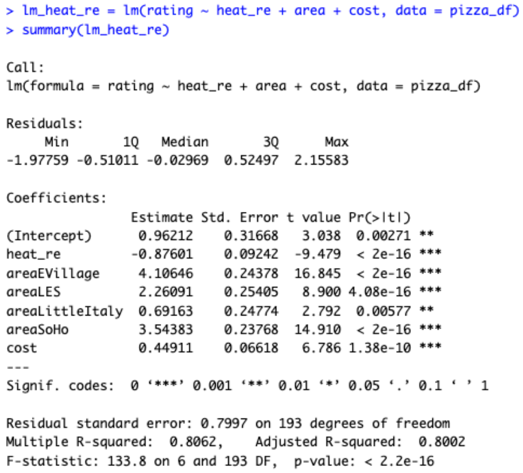
f. **heat_re**, **area** and **cost** as the predictor variables.

Assume that coal-baked pizzas produce the highest ratings, followed by using wood, and then gas, compare the two models. It is more reasonable to use dummy (indicator) variables in model fitting (as in 1b.), why? Justify your answer by comparing the interpretations of the regression coefficients of **heat** and **heat_re**.

Ans：因為模型可以更好的捕捉到 heat 的影響，不會將影響歸因於 heat 的順序，可以看到在 lm_heat 中”heatGas”和”heatWood”的係數分別是-1.59555和-0.45753，分別表示使用”Gas”和”Wood”的烘烤方式對於 rating 的影響，但在 lm_heat_re 模型中只有”heat_re”係數為-0.87601，表示使用”Gas”或”Wood”烘烤方式對 rating 的影響，可以看出 lm_heat模型提供更多的細節與資訊。



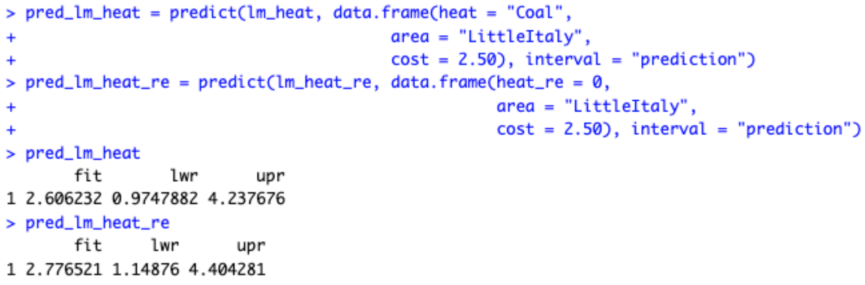
```
> lm_heat = lm(rating ~ heat + area + cost, data = pizza_df)
> summary(lm_heat)

Call:
lm(formula = rating ~ heat + area + cost, data = pizza_df)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.98864 -0.52516  0.00599  0.51428  1.92332 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      0.72260    0.34461   2.097  0.03731 *  
heatGas         -1.59555    0.20526  -7.773 4.52e-13 ***
heatWood        -0.45753    0.26056  -1.756  0.08069 .  
areaEVillage     4.17970    0.24628  16.971  < 2e-16 ***
areaLES          2.37294    0.26106   9.089  < 2e-16 ***
areaLittleItaly  0.78700    0.25268   3.115  0.00212 ** 
areaSoHo         3.65362    0.24498  14.914  < 2e-16 ***
cost             0.43865    0.06613   6.633 3.26e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7957 on 192 degrees of freedom
Multiple R-squared:  0.8092,	Adjusted R-squared:  0.8022 
F-statistic: 116.3 on 7 and 192 DF,  p-value: < 2.2e-16
```



```
> lm_heat_re = lm(rating ~ heat_re + area + cost, data = pizza_df)
> summary(lm_heat_re)

Call:
lm(formula = rating ~ heat_re + area + cost, data = pizza_df)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.97759 -0.51011 -0.02969  0.52497  2.15583 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      0.96212    0.31668   3.038  0.00271 ** 
heat_re         -0.87601    0.09242  -9.479  < 2e-16 ***
areaEVillage     4.10646    0.24378  16.845  < 2e-16 ***
areaLES          2.26091    0.25405   8.900 4.08e-16 ***
areaLittleItaly  0.69163    0.24774   2.792  0.00577 ** 
areaSoHo         3.54383    0.23768  14.910  < 2e-16 ***
cost             0.44911    0.06618   6.786 1.38e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7997 on 193 degrees of freedom
Multiple R-squared:  0.8062,	Adjusted R-squared:  0.8002 
F-statistic: 133.8 on 6 and 193 DF,  p-value: < 2.2e-16
```

Then, predict the rating for a coal baked pizza that costs $2.50 per slice in LittleItaly and find the corresponding prediction interval using both of the models built in 3a. and 3b. *[hint: use **predict()**]*

Ans :



```
> pred_lm_heat = predict(lm_heat, data.frame(heat = "Coal",
+                                            area = "LittleItaly",
+                                            cost = 2.50), interval = "prediction")
> pred_lm_heat_re = predict(lm_heat_re, data.frame(heat_re = 0,
+                                                  area = "LittleItaly",
+                                                  cost = 2.50), interval = "prediction")
> pred_lm_heat
       fit       lwr      upr
1 2.606232 0.9747882 4.237676
> pred_lm_heat_re
       fit     lwr      upr
1 2.776521 1.14876 4.404281
```

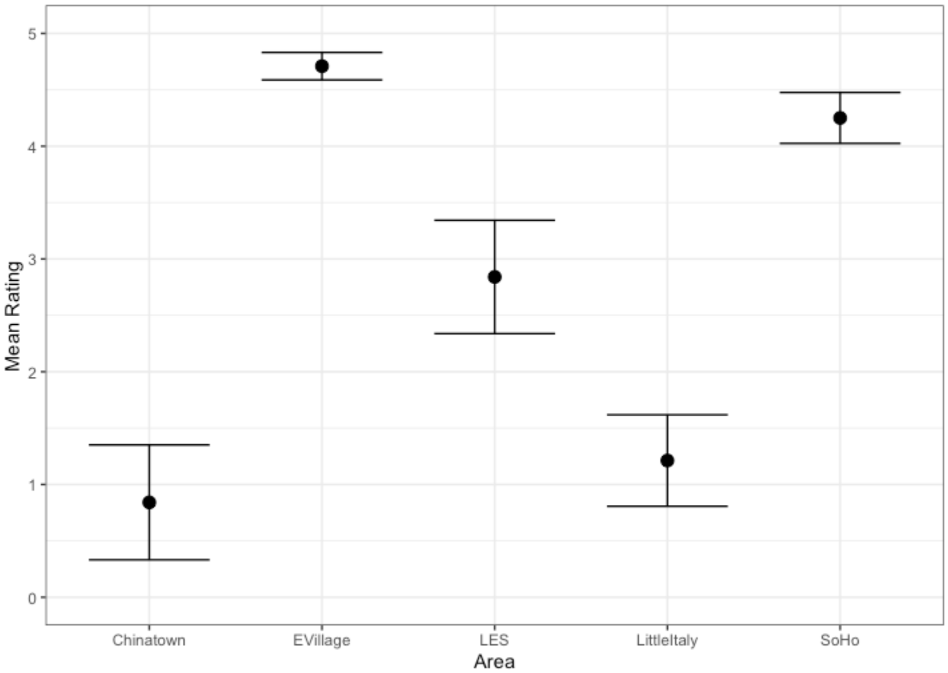
(“heat” 為 categorical variable, “heat_re” 則是 numerical. 在這題用兩個方法跑迴歸, 再比較兩者結果.)

III. Construct the 95% t-based confidence intervals for the mean rating for each pizzeria location (**area**). Plot **all** of the intervals in a single plot and briefly comment the results. (*Hint: you could make use of **plot(), lines()** and **points()** **OR** search online*[1] *for some ways to plot confidence intervals.*)

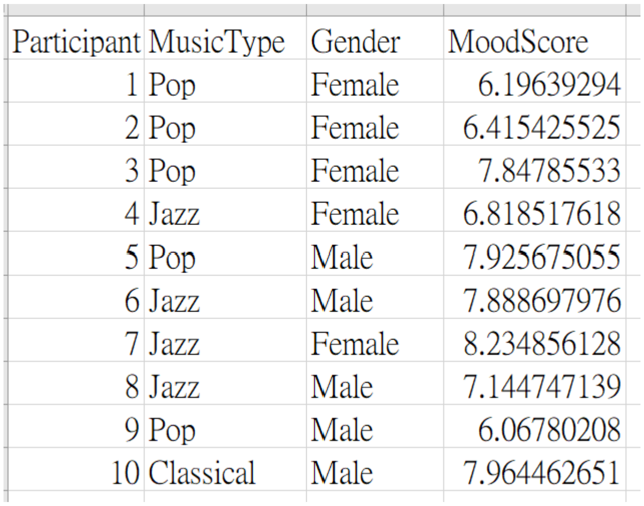
練習畫信賴區間 (後來的 project 可以使用類似方法呈現資料的比較)

Ans : 從圖中可以看到 EVillage 和 SoHo 地區的 mean rating 較高但 confidence interval 較窄，表示這些地區的 rating 較穩定，而其他地區的 mean rating 較低且 confidence interval 較寬，表示這些地區的 rating 較不穩定。

[1] http://stackoverflow.com/questions/14069629/plotting-confidence-intervals



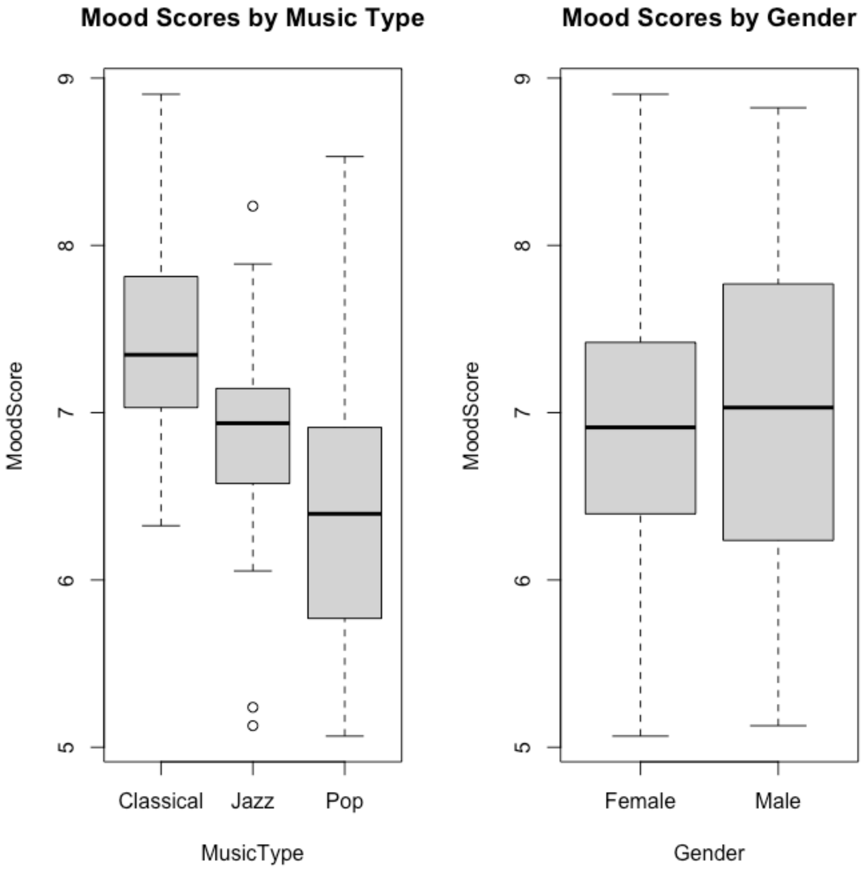
2. Suppose we are interested in studying the effect of different types of music on people's moods. We collect data on 60 participants and record their mood score (out of 10) after listening to one of three types of music: classical, jazz, or pop. In the file "mood.csv" The data look like



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

Let Y be "MoodScore" and "MusicType" and "Gender" be explanatory variables.

a. Draw side-by-side boxplots to compare the distribution of mood scores based on the music type. Also side-by-side boxplots to compare the distribution of mood scores based on gender.

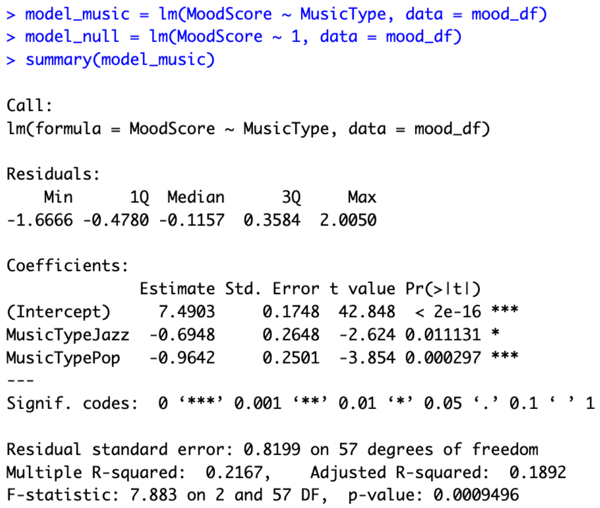


b. First, test whether the type of music has a significant effect on mood scores while ignoring gender in the model. (Hint: Use a partial F test. Write down the regression model expression used in the analysis.)

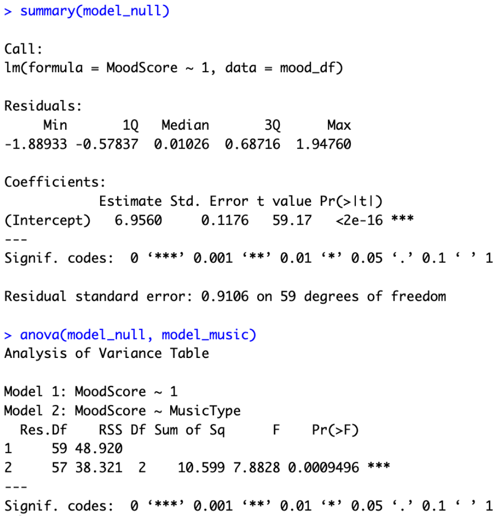
Ans：根據結果可以得知有 MusicType 解釋變數模型比沒有解釋變數的模型更好的解釋資料，因為 partial F test 的 p-value 非常小，而且 Sum of Sq 為 10.599 很大表示兩個模型之間有顯著的差異。而 MusicType 中 Classical 是最有顯著影響接著是 Pop 最後才是 Jazz。

Model_music expression : $\text{MoodScore} = 7.4903 + (-0.6948) * MusicTypeJazz + (-0.9642) * \ MusicTypePop$

Model_null expression : $\text{MoodScore} = 6.9560$



```
> model_music = lm(MoodScore ~ MusicType, data = mood_df)
> model_null = lm(MoodScore ~ 1, data = mood_df)
> summary(model_music)

Call:
lm(formula = MoodScore ~ MusicType, data = mood_df)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.6666 -0.4780 -0.1157  0.3584  2.0050 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)     7.4903     0.1748  42.848  < 2e-16 ***
MusicTypeJazz  -0.6948     0.2648  -2.624 0.011131 *  
MusicTypePop   -0.9642     0.2501  -3.854 0.000297 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8199 on 57 degrees of freedom
Multiple R-squared:  0.2167,	Adjusted R-squared:  0.1892 
F-statistic: 7.883 on 2 and 57 DF,  p-value: 0.0009496
```



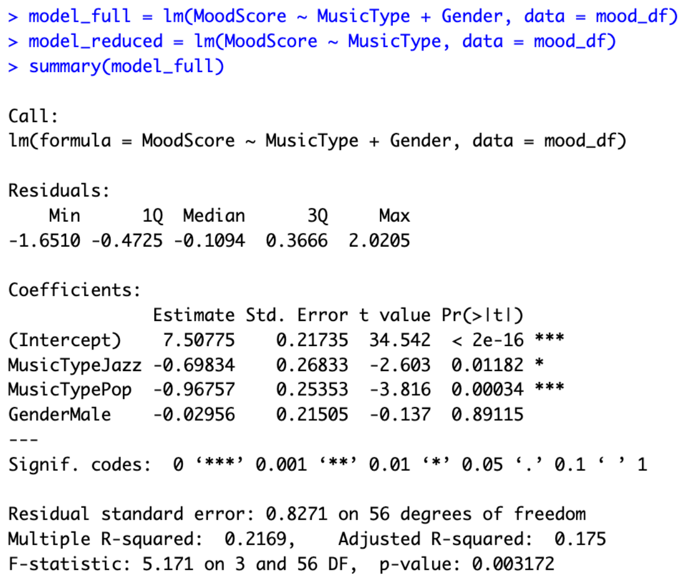
```
> summary(model_null)

Call:
lm(formula = MoodScore ~ 1, data = mood_df)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.88933 -0.57837  0.01026  0.68716  1.94760 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   6.9560     0.1176   59.17   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9106 on 59 degrees of freedom

> anova(model_null, model_music)
Analysis of Variance Table

Model 1: MoodScore ~ 1
Model 2: MoodScore ~ MusicType
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1     59 48.920                                  
2     57 38.321  2    10.599 7.8828 0.0009496 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

c. Then, test whether the type of music has a significant effect on mood scores while including gender in the model. (Hint: Use a partial F test. Write down the regression model expression used in the analysis.)

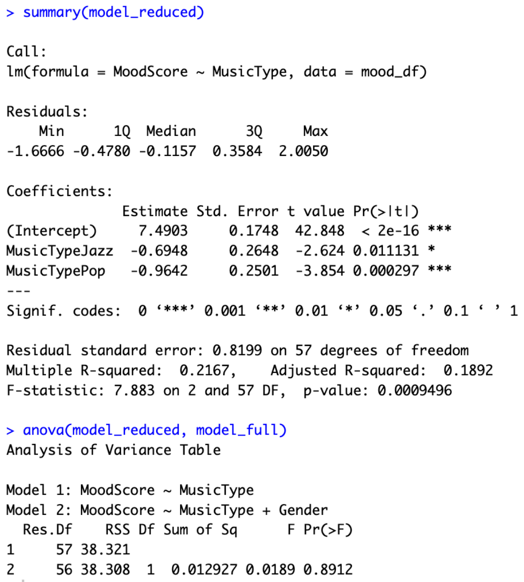
Ans : 由 partial F test 中的 p-value 為 0.8912 大於 0.05 無法拒絕虛無假設，所以在加入 Gender 的情況下，對 MoodScore 並無顯著影響，而且 Sum of Sq 為 0.012927 非常小表示兩個模型之間沒有太大的顯著差異。Classical 是最有顯著影響接著是 Pop 最後才是 Jazz。

Model_full expression : $ModelScore = 7.50775 + (-0.69834) * MusicTypeJazz + (-0.96757) * \ MusicTypePop + (-0.02956) * GenderMale$

Model_reduced expression : $ModelScore = \ 7.4903 + (-0.6948) * MusicTypeJazz + (-0.9642) * \ MusicTypePop$



```
> model_full = lm(MoodScore ~ MusicType + Gender, data = mood_df)
> model_reduced = lm(MoodScore ~ MusicType, data = mood_df)
> summary(model_full)

Call:
lm(formula = MoodScore ~ MusicType + Gender, data = mood_df)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.6510 -0.4725 -0.1094  0.3666  2.0205 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    7.50775    0.21735  34.542  < 2e-16 ***
MusicTypeJazz -0.69834    0.26833  -2.603  0.01182 *  
MusicTypePop  -0.96757    0.25353  -3.816  0.00034 ***
GenderMale    -0.02956    0.21505  -0.137  0.89115    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8271 on 56 degrees of freedom
Multiple R-squared:  0.2169,	Adjusted R-squared:  0.175 
F-statistic: 5.171 on 3 and 56 DF,  p-value: 0.003172
```

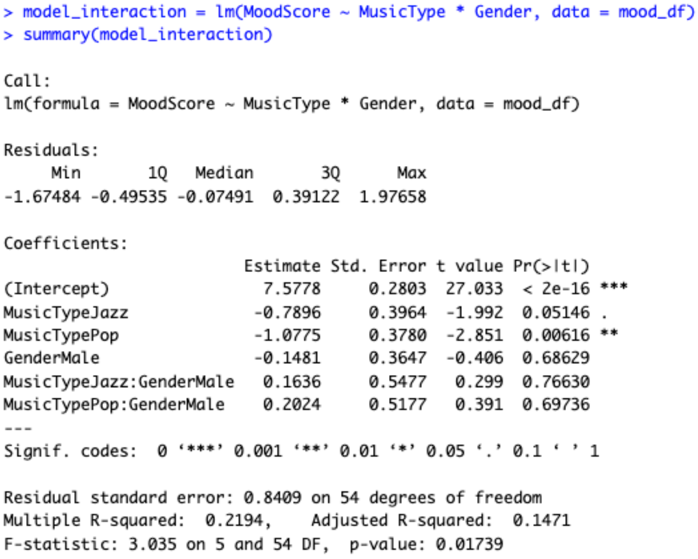


```
> summary(model_reduced)

Call:
lm(formula = MoodScore ~ MusicType, data = mood_df)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.6666 -0.4780 -0.1157  0.3584  2.0050 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)     7.4903     0.1748  42.848  < 2e-16 ***
MusicTypeJazz  -0.6948     0.2648  -2.624 0.011131 *  
MusicTypePop   -0.9642     0.2501  -3.854 0.000297 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8199 on 57 degrees of freedom
Multiple R-squared:  0.2167,	Adjusted R-squared:  0.1892 
F-statistic: 7.883 on 2 and 57 DF,  p-value: 0.0009496

> anova(model_reduced, model_full)
Analysis of Variance Table

Model 1: MoodScore ~ MusicType
Model 2: MoodScore ~ MusicType + Gender
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     57 38.321                           
2     56 38.308  1  0.012927 0.0189 0.8912
```

d. Finally, test whether there exists an interaction effect between gender and the type of music on mood scores. (Hint: Use a t test. Write down the regression model expression used in the analysis.)

Ans：由 MusicTypeJazz:GenderMale 和 MusicTypePop:GenderMale 的 p-value 皆大於 0.05，表示無法拒絕虛無假設，代表 MusicType 和 Gender 之間不存在交互作用。

Model_interaction expression : $MoodScore = 7.5778 + (-0.7896) * Jazz + (-1.0775) * Pop + (-0.1481) * Male + 0.1636 * Jazz: Male + 0.2024 * Pop: Male$



```
> model_interaction = lm(MoodScore ~ MusicType * Gender, data = mood_df)
> summary(model_interaction)

Call:
lm(formula = MoodScore ~ MusicType * Gender, data = mood_df)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.67484 -0.49535 -0.07491  0.39122  1.97658 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)                7.5778     0.2803  27.033  < 2e-16 ***
MusicTypeJazz             -0.7896     0.3964  -1.992  0.05146 .  
MusicTypePop              -1.0775     0.3780  -2.851  0.00616 ** 
GenderMale                -0.1481     0.3647  -0.406  0.68629    
MusicTypeJazz:GenderMale   0.1636     0.5477   0.299  0.76630    
MusicTypePop:GenderMale    0.2024     0.5177   0.391  0.69736    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8409 on 54 degrees of freedom
Multiple R-squared:  0.2194,	Adjusted R-squared:  0.1471 
F-statistic: 3.035 on 5 and 54 DF,  p-value: 0.01739
```