Assignment Semester 1 2022

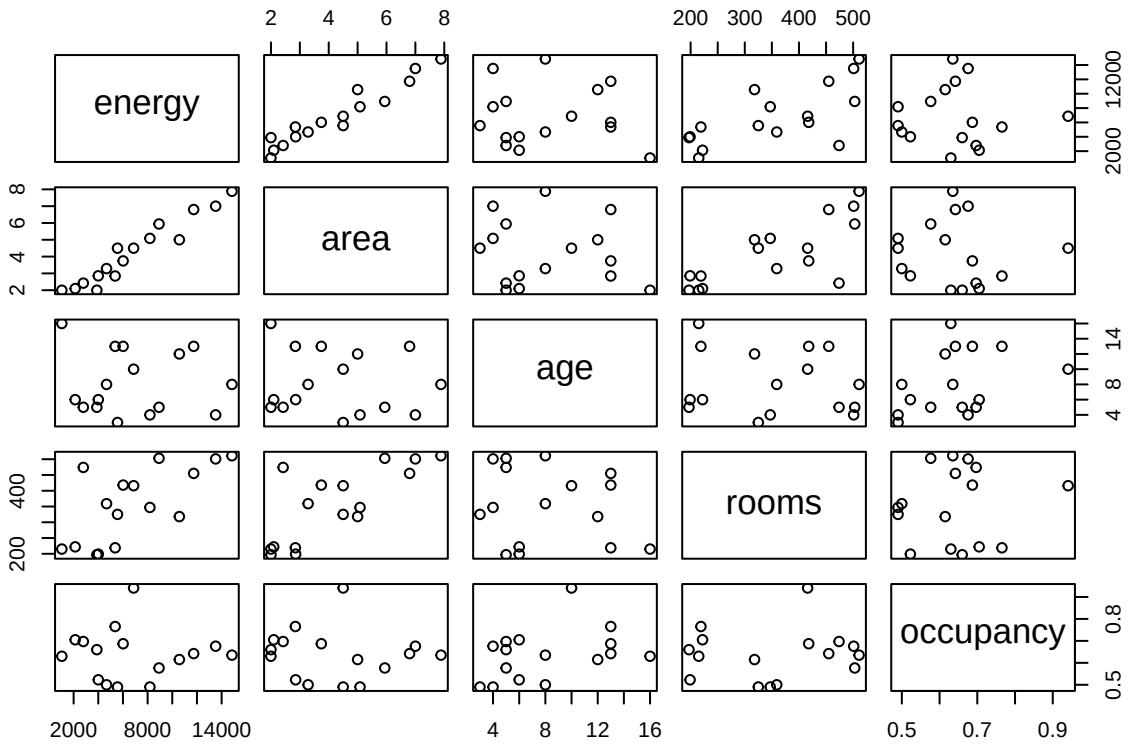
Question 1

a. Produce a plot and a correlation matrix of the data. Comment on possible relationships between the response and predictors and relationships between the predictors themselves.

```
hotel <- read.table("hotel2022 (1).dat", header = T)
head(hotel)
```

```
##     energy   area age rooms occupancy
## 1 1045.555 1.9979  16   215    0.6300
## 2 2126.199 2.0962   6   222    0.7050
## 3 2785.958 2.4212   5   474    0.6970
## 4 5558.123 4.5000   3   325    0.4900
## 5 4001.213 2.8548   6   199    0.5227
## 6 4669.758 3.2865   8   359    0.5000
```

```
pairs(hotel)
```



```
cor(hotel)
```

```
##                energy       area         age       rooms   occupancy
## energy      1.00000000  0.9631684 -0.05707197  0.68265406 -0.02022606
## area        0.96316841  1.0000000 -0.10787993  0.76068692 -0.08924440
## age        -0.05707197 -0.1078799  1.00000000 -0.16142400  0.36195114
## rooms       0.68265406  0.7606869 -0.16142400  1.00000000  0.09944548
## occupancy -0.02022606 -0.0892444  0.36195114  0.09944548  1.00000000
```

From above plot and correlation matrix, I observed that the from all predictors are has strongest correlation with energy. The correlation value between these two variables is 0.96. The predictor rooms also has stronger relation with dependent variable energy. The correlation value between these two variables is 0.68. From all predictors area and rooms has strongest relation. The correlation value between these two variables is 0.76.

- **Fit a model using all the predictors to explain the energy response.**

```
mod <- lm(energy ~ ., hotel)
summary(mod)
```

```
##
## Call:
## lm(formula = energy ~ ., data = hotel)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1577.1  -720.8   118.7   608.6  1809.4
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3197.279   1871.533  -1.708    0.116
## area         2331.116    250.919   9.290 1.53e-06 ***
## age             2.358     80.841   0.029    0.977
## rooms          -5.383      4.168  -1.291    0.223
## occupancy    3234.553   2928.605   1.104    0.293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1154 on 11 degrees of freedom
## Multiple R-squared:  0.9416, Adjusted R-squared:  0.9203
## F-statistic:  44.3 on 4 and 11 DF,  p-value: 1.019e-06
```

- **Using the full model, estimate the impact each hectare of hotel area has on the energy efficiency of the hotel. Do this by producing a 95% confidence interval that quantifies the change in energy consumption for each extra hectare of hotel area and comment.**

```
confint(mod)
```

```
##                   2.5 %      97.5 %
```

```
## (Intercept) -7316.49578  921.937963
## area         1778.84733 2883.385146
## age          -175.57318  180.288690
## rooms         -14.55762    3.791228
## occupancy   -3211.26348 9680.370316
```

The 95% confidence interval for are is [1778.84, 2883.38] which means we are 95% confident that if we hold other predictors constant and increase the value of are by one hectare the energy consumption will be between 1778.84 to 2883.38.

c Conduct an F-test for the overall regression i.e. is there any relationship between the response and the predictors. In your answer:

- **Write down the mathematical multiple regression model for this situation, defining all appropriate parameters.**

```
summary(mod)
```

```
##
## Call:
## lm(formula = energy ~ ., data = hotel)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1577.1  -720.8   118.7   608.6  1809.4
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3197.279   1871.533  -1.708    0.116
## area         2331.116    250.919   9.290 1.53e-06 ***
## age             2.358     80.841   0.029    0.977
## rooms          -5.383      4.168  -1.291    0.223
## occupancy    3234.553   2928.605   1.104    0.293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1154 on 11 degrees of freedom
## Multiple R-squared:  0.9416, Adjusted R-squared:  0.9203
## F-statistic:  44.3 on 4 and 11 DF,  p-value: 1.019e-06
```

The equation for multiple regression model is:

$$energy = -3197.279 + 2331.11 * area + 2.358 * age - 5.383 * rooms + 3234.554 * occupancy$$

• Write down the Hypotheses for the Overall ANOVA test of multiple regression. The null and alternative hypotheses for the overall anova test of multiple regression are shown below: H0: There is no significant relation between the response variable and predictors. There is a significant relation between atleast one of the predictors and response variable. The significance level alpha = 0.05.

```
anova(mod)
```

- **Produce an ANOVA table for the overall multiple regression model (One combined regression SS source is sufficient).**

```
## Analysis of Variance Table
##
## Response: energy
##           Df    Sum Sq   Mean Sq  F value    Pr(>F)
## area       1 232665641 232665641 174.5879 4.297e-08 ***
## age        1    556602    556602   0.4177    0.5314
## rooms      1   1293045   1293045   0.9703    0.3458
## occupancy  1   1625643   1625643   1.2199    0.2930
## Residuals 11  14659219   1332656
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(mod)
```

```
##
## Call:
## lm(formula = energy ~ ., data = hotel)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1577.1  -720.8   118.7   608.6  1809.4
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3197.279   1871.533  -1.708    0.116
## area         2331.116    250.919   9.290 1.53e-06 ***
## age             2.358     80.841   0.029    0.977
## rooms          -5.383      4.168  -1.291    0.223
## occupancy    3234.553   2928.605   1.104    0.293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1154 on 11 degrees of freedom
## Multiple R-squared:  0.9416, Adjusted R-squared:  0.9203
## F-statistic:  44.3 on 4 and 11 DF,  p-value: 1.019e-06
```

- **Compute the F statistic for this test.**

The value of F-statistic is 44.3.

- **State the Null distribution for the test statistic.**

The null distribution for test statistic is the multiple regression model is not statistically significant.

- **Compute the P-Value** The p-value is 1.019e-06.

• State your conclusion (both statistical conclusion and contextual conclusion). Since the p-value is less than significance level alpha, so I reject the null hypothesis and conclude that there is a significant relation between at least one predictor and response variable so the model is statistically significant.

###d. [10 marks] Validate the full model and comment on whether the full regression model is appropriate to explain the energy efficiency of the hotels. There is only one predictor energy, that has p-value of coefficient less than significance level. All other predictors don't have statistically significant relation with dependent variable so the full regression model is not appropriate at all.

e. [2 marks] Find the R2 and comment on what it means in the context of this dataset.

```
summary(mod)
```

```
##
## Call:
## lm(formula = energy ~ ., data = hotel)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1577.1  -720.8   118.7   608.6  1809.4
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3197.279   1871.533  -1.708    0.116
## area         2331.116    250.919   9.290 1.53e-06 ***
## age             2.358     80.841   0.029    0.977
## rooms          -5.383      4.168  -1.291    0.223
## occupancy    3234.553   2928.605   1.104    0.293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1154 on 11 degrees of freedom
## Multiple R-squared:  0.9416, Adjusted R-squared:  0.9203
## F-statistic:  44.3 on 4 and 11 DF,  p-value: 1.019e-06
```

The R-squared value is 0.9416 which means 94.16% of variation in energy consumption is explained by all predictors.

```
null <- lm(energy ~ 1, data = hotel)
full <- lm(energy ~., data = hotel)
best_mod <- step(null, direction='forward', scope=formula(full), trace=0)
summary(best_mod)
```

f. [3 marks] Using model selection procedures discussed in the course, find the best multiple regression model that explains the data. State the final fitted regression model.

```
##
## Call:
## lm(formula = energy ~ area, data = hotel)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1843.6  -447.1  -275.4   580.8  2141.0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1874.6      712.5  -2.631   0.0198 *
## area          2061.4      153.8  13.402 2.24e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1138 on 14 degrees of freedom
## Multiple R-squared:  0.9277, Adjusted R-squared:  0.9225
## F-statistic: 179.6 on 1 and 14 DF,  p-value: 2.237e-09
```

The final fitted regression model has only one predictor that is area. The regression equation is:

$$energy = -1874.6 + 2061.4 * area$$

g. Comment on the R2 and adjusted R2 in the full and final model you chose in part d. In particular explain why those goodness of fitness measures change but not in the same way. The R-squared value of final model is 0.9277 which means 92.77% of variation in the dependent variable is explained by the predictors. The R-squared value is bit less as compare to the full model which is understandable as that model has more predictor. Since the final model has only one predictor that's why the fitness measures changed.

Question 2

```
#reading data
movies <- read.table("movie (2).dat", header = T)
head(movies)
```

a. [2 marks] For this study, is the design balanced or unbalanced? Explain why.

```
##   Gender  Genre Score
## 1      F Action     1
## 2      F Action     1
## 3      F Action     1
## 4      F Action     1
## 5      F Action     2
## 6      F Action     2
```

The balance design's are one that have equal number of observations for all possible level of combinations. For checking the number of observations for all possible level of combinations, I used table function.

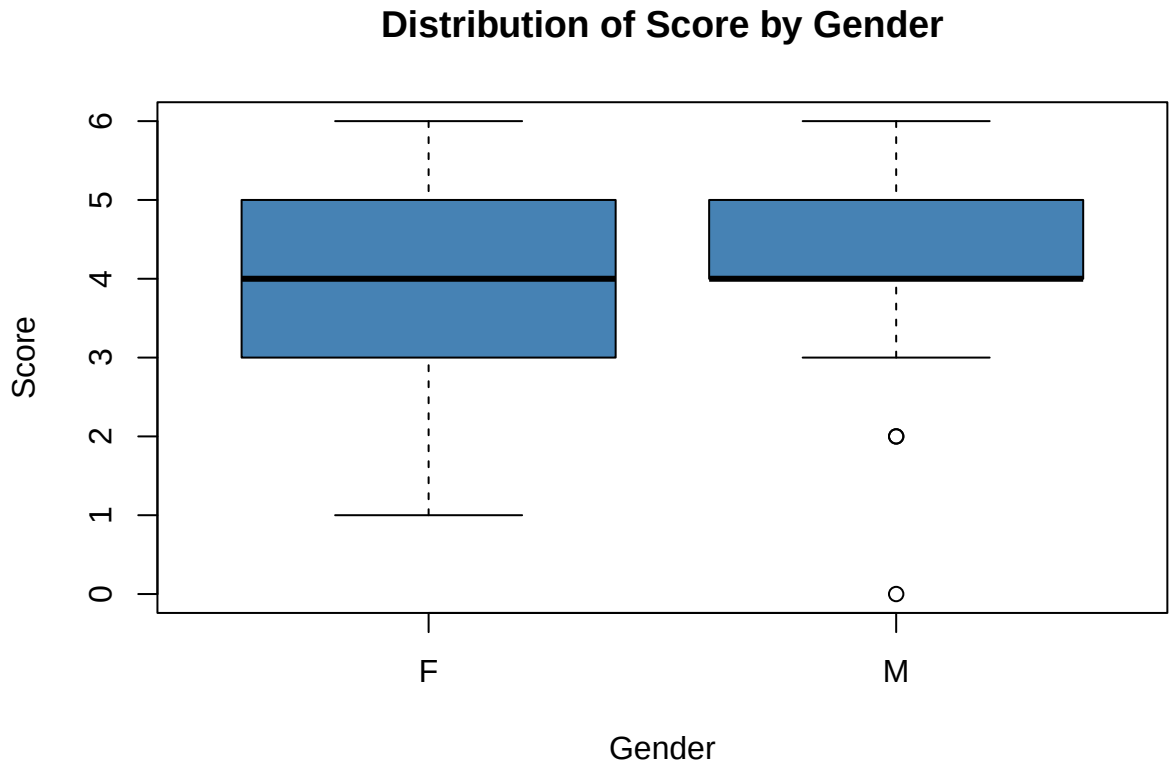
```
table(movies$Gender, movies$Genre)
```

```
##
##     Action Comedy Drama
##   F     39     33    22
##   M     14     10    19
```

Since there is a difference in the number of observations for different level of combinations so the design is not balanced and it is unbalanced.

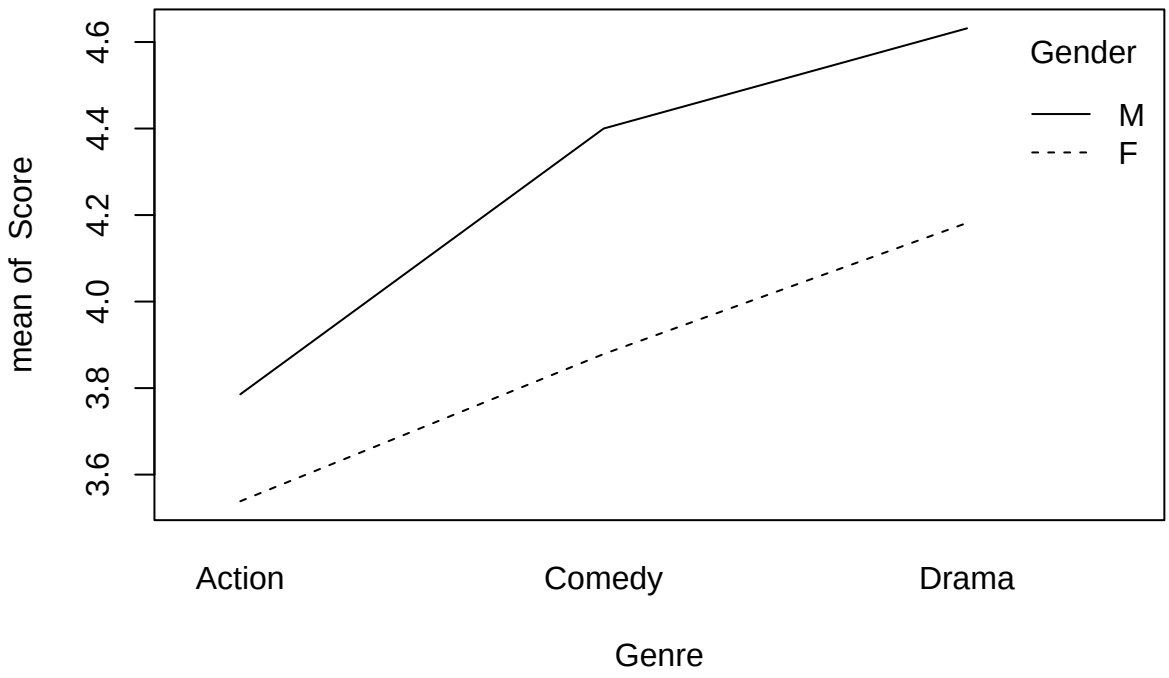
b. Construct two different preliminary graphs that investigate different features of the data and comment.

```
boxplot(Score ~ Gender, movies, col = "steelblue", main = "Distribution of Score by Gender")
```



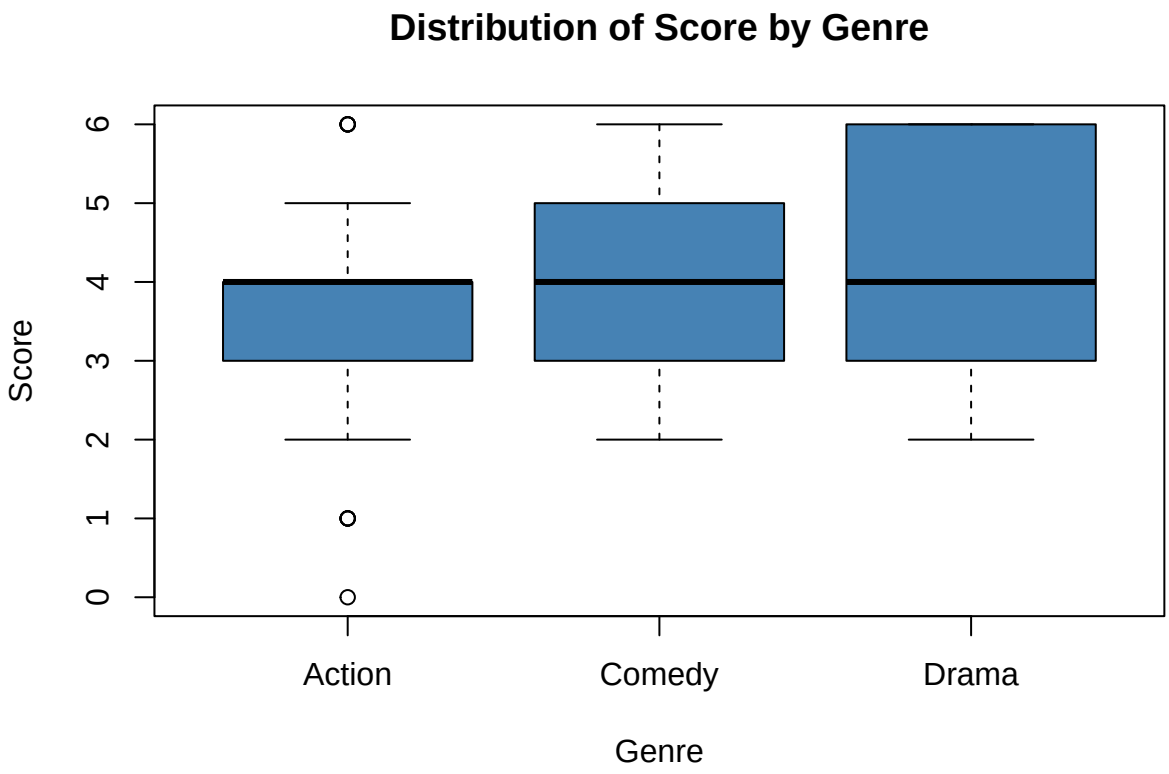
From above plot, it is pretty evident that males have higher average value of score as compare to female.

```
#ineraction plot
with(movies, interaction.plot(Genre, Gender, Score))
```



From above plot it seems that Female tends to have lower mean score as compare to the Male. Action seems to have lowest mean score and Drame seems to have highest mean score

```
boxplot(Score ~ Genre, movies, col = "steelblue", main = "Distribution of Score by Genre")
```



From above plot, it is pretty evident that almost all three genres have equal median value of scores. Only Action genre has outliers with low and high value of Scores.

```
summary(lm(Score ~ ., data = movies))
```

c. [4 marks] Write down the full mathematical model for this situation, defining all appropriate parameters.

```
##
## Call:
## lm(formula = Score ~ ., data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8946 -0.9081  0.0919  1.0919  2.5006
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.4994     0.1930  18.134   <2e-16 ***
## GenderM       0.3952     0.2489   1.588   0.1147
## GenreComedy   0.4087     0.2712   1.507   0.1342
## GenreDrama    0.7077     0.2792   2.535   0.0124 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.321 on 133 degrees of freedom
## Multiple R-squared:  0.07488,    Adjusted R-squared:  0.05401
## F-statistic: 3.588 on 3 and 133 DF,  p-value: 0.01552
```

The mathematical equation in this case will be:

$$Scores = b0 + b1 * GenderM + b2 * GenreComedy + b3 * GenreDrama$$

#d. [15 marks] Analyse the data to study the effect of Gender and Genre on the brand recall Score. These conclusions are only required to be at the qualitative level and can be based off the outcomes of the hypothesis tests you conducted in this part and the preliminary plots in part b. You do not need to statistically examine the multiple comparisons between contrasts and interactions. Remember to # • state the null and alternative hypothesis for each test, and # • check assumptions.

The null and alternative hypotheses are given below: H0:There is no relation between Score and any of predictors. Ha: There is a significant relation between Score and at least one of Gender or Genre. The significant level alpha = 0.05.

```
#first model
mod1 <- lm(Score ~ Gender + Genre, movies)
anova(mod1)
```

```
## Analysis of Variance Table
##
## Response: Score
##            Df  Sum Sq Mean Sq F value  Pr(>F)
## Gender      1   7.195  7.1946  4.1238 0.04428 *
## Genre       2  11.587  5.7934  3.3207 0.03915 *
## Residuals 133 232.036  1.7446
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value for both Gender and Genre are less than significant level 0.05 so both have significant relation with Score.

```
#second  model
mod2 <- lm(Score ~ Genre + Gender, movies)
anova(mod2)
```

```
## Analysis of Variance Table
##
## Response: Score
##            Df  Sum Sq Mean Sq F value  Pr(>F)
## Genre       2  14.382  7.1911  4.1218 0.01833 *
## Gender      1   4.399  4.3993  2.5216 0.11467
## Residuals 133 232.036  1.7446
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From this model the p-value of Genre is less than 0.05 so Genre has significant relation with Score and Gender has p-value greater than 0.05 which means Gender hasn't significant relation with score.

```r
#model with interaction term
mod3 <- lm(Score ~ Genre*Gender, movies)
anova(mod3)
```

```
## Analysis of Variance Table
##
## Response: Score
##               Df  Sum Sq Mean Sq F value  Pr(>F)
## Genre          2  14.382  7.1911  4.0665 0.01935 *
## Gender         1   4.399  4.3993  2.4877 0.11715
## Genre:Gender   2   0.378  0.1889  0.1068 0.89879
## Residuals    131 231.658  1.7684
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the interaction terms p-value is greater than significant level alpha = 0.05 so the interaction term is not statistically significant.