47/70

Assignment [redacted]

[redacted]

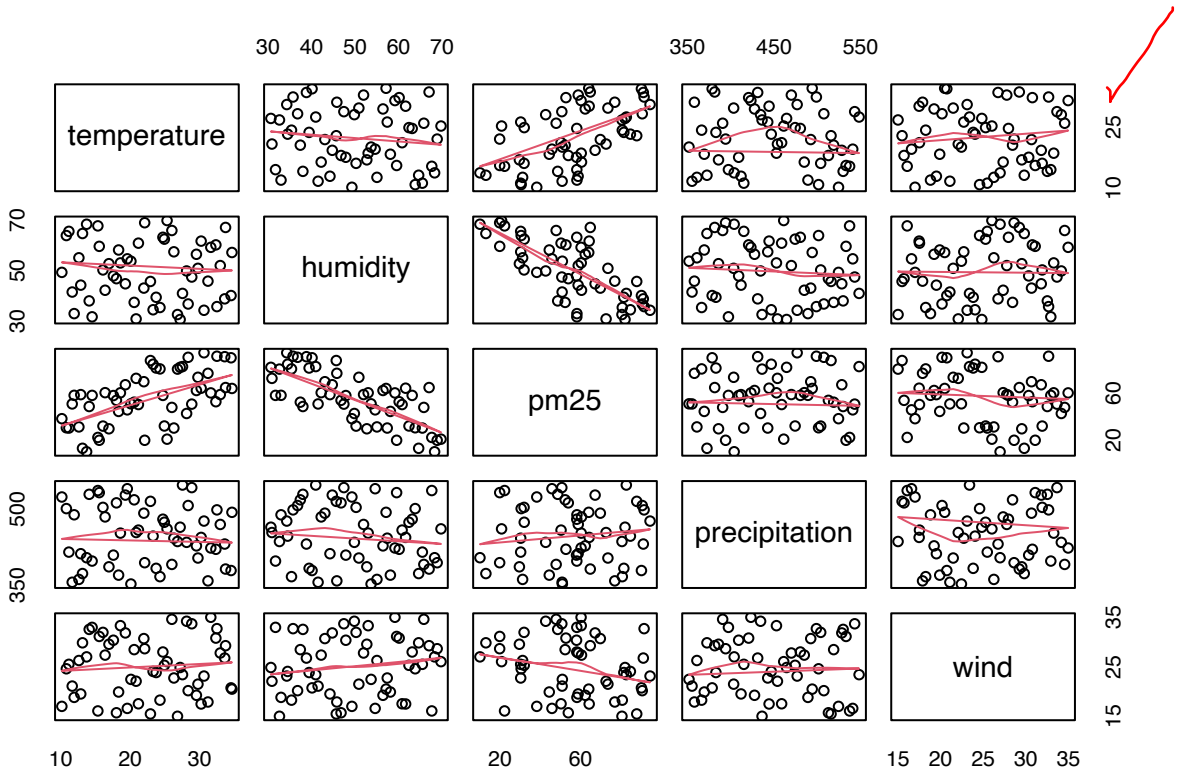
[redacted]

Question 1 28/45

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. 6/7

```
pairs(pm25, panel = panel.smooth)
```

Where is the read file?



Comments: We can observe a moderate correlation between PM2.5 and variable humidity and variable temperature. You should elaborate more on the various relationships.

```
cor(pm25)
```

```
##                temperature    humidity        pm25 precipitation        wind
## temperature     1.00000000 -0.07264891  0.57191961   -0.05050014  0.02861166
## humidity       -0.07264891  1.00000000 -0.71965591   -0.13550607  0.12406351
## pm25            0.57191961 -0.71965591  1.00000000    0.03759033 -0.21866823
## precipitation -0.05050014 -0.13550607  0.03759033    1.00000000 -0.01525977
## wind           0.02861166  0.12406351 -0.21866823   -0.01525977  1.00000000
```

Comments: Based on the correlation matrix, there is no visible correlation between four predictors. The correlation coefficients between temperature and the other three predictors approximate 0.00.

This could do with more detail.

b) 6/6

- Fit a model using all the predictors to explain the pm25 response.

```
pm25.lm = lm(pm25 ~ temperature + humidity + precipitation + wind, data = pm25)
summary(pm25.lm)
```

```
##
## Call:
## lm(formula = pm25 ~ temperature + humidity + precipitation +
##     wind, data = pm25)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.759  -6.804  -1.649   6.857  20.975
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   102.72259   14.71953   6.979 5.88e-09 ***
## temperature     1.62142    0.18762   8.642 1.46e-11 ***
## humidity       -1.27742    0.11854 -10.776 9.49e-15 ***
## precipitation  -0.01091    0.02350  -0.464   0.6444
## wind           -0.58016    0.23405  -2.479   0.0165 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.06 on 51 degrees of freedom
## Multiple R-squared:  0.8127, Adjusted R-squared:  0.7981
## F-statistic: 55.34 on 4 and 51 DF,  p-value: < 2.2e-16
```

Regression Model: $\hat{PM2.5} = 102.72259 + 1.62142 \times \text{temperature} - 1.27742 \times \text{humidity} - 0.01091 \times \text{precipitation} - 0.58016 \times \text{wind} + \varepsilon$

- Using the full model, estimate the impact of humidity on PM2.5 concentration. Do this by producing a 95% confidence interval that quantifies the change in PM2.5 concentration for each extra percentage of relative humidity and comment.

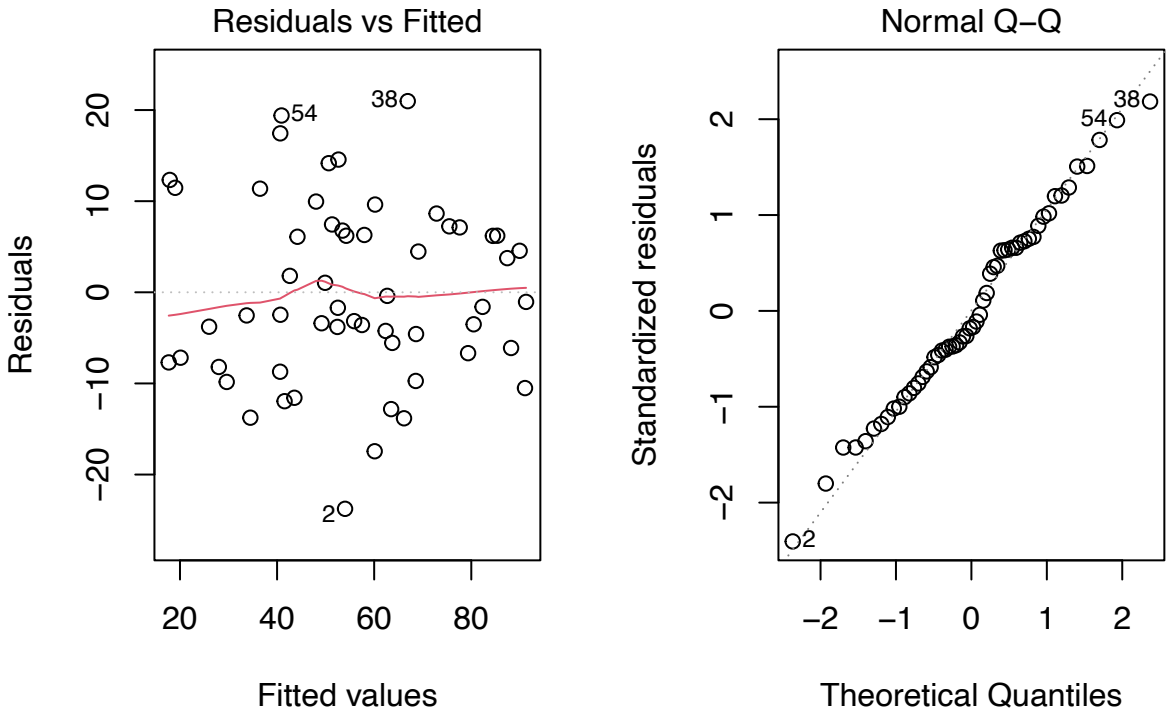
$\hat{\beta_{humidity}} \pm t \times s.e.(\hat{\beta_{humidity}}) = -1.27742 \pm 2.007584 \times 0.11854 = (-1.515399, -1.039441)$

While you are correct, please include more detail.

Comments: For each extra percentage increase of humidity, we would expect a decrease in PM2.5 concentration between 1.039441 and 1.515399.

c) Conduct an F-test for the overall regression 11/14

```
par(mfrow = c(1, 2))
plot(pm25.lm, which = 1:2)
```



Comments: There is no particular pattern in the residuals vs fitted plot, and the linear trend is visible in the Normal Q-Q plot, which indicates that the residuals are close to normally distributed. Therefore, they satisfy the constant variance and normality assumptions. True but unneeded

Multiple Regression Model:

$$\hat{PM2.5} = \beta_0 + \beta_1 \times temperature + \beta_2 \times humidity + \beta_3 \times precipitation + \beta_4 \times wind + \varepsilon \ , \ \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

Where:
PM2.5 is the dependent variable.
β_0 represents the intercept.
$\beta_1, \beta_2, \beta_3, \beta_4$ represent the coefficients of independent variables.
temperature, humidity, precipitation, wind are the independent variables.
ε represents the error term or residuals.

Hypotheses:

$H_0 : \beta_{temperature} = \beta_{humidity} = \beta_{precipitation} = \beta_{wind} = 0$

H_1 : not all $\beta_i = 0$

```
anova_pm25 <- anova(pm25.lm)
print(anova_pm25)
```

```
## Analysis of Variance Table
##
## Response: pm25
##               Df  Sum Sq Mean Sq  F value    Pr(>F)
## temperature    1  9014.4  9014.4  89.0853 8.908e-13 ***
## humidity       1 12739.7 12739.7 125.9013 2.200e-15 ***
## precipitation  1    22.7    22.7   0.2247   0.63752
## wind           1   621.7   621.7   6.1442   0.01653 *
## Residuals     51  5160.6   101.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test Statistic: $F_{obs} = \frac{Reg.M.S.}{Res.M.S.} = \frac{5599.625}{101.1882} = 55.33872$ Where is the ANOVA table? Sum Sq???

Null distribution: If H_0 is true, F_{obs} follows a $F_{4,51}$ distribution

P-Value: $P(F_{4,51} \geq 55.34) = 6.082082e-18 < 0.05$

Conclusion: Because the $P - Value$ is equal to 6.082082e-18, which is much smaller than the significance level of 5%, therefore, we do not have enough evidence to reject the null hypothesis. Wrong conclusion, you do reject

d) Validate the full model and comment on whether the full regression model is appropriate to explain the PM2.5 concentration at various test locations 0/10

```
anova_pm25 <- anova(pm25.lm)
print(anova_pm25)
```

```
## Analysis of Variance Table
##
## Response: pm25
##               Df  Sum Sq Mean Sq  F value    Pr(>F)
## temperature    1  9014.4  9014.4  89.0853 8.908e-13 ***
## humidity       1 12739.7 12739.7 125.9013 2.200e-15 ***
## precipitation  1    22.7    22.7   0.2247   0.63752
## wind           1   621.7   621.7   6.1442   0.01653 *
## Residuals     51  5160.6   101.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comments: The full model has the precipitation predictor, which is insignificant (P-Value = 0.63752 > 0.05), therefore, it is inappropriate to use the full model to explain the PM2.5 concentration at various test locations. It is suggested to proceed with a new regression model without the variable precipitation.

You were asked to validate the FULL model.

```r
pm25.lm2 = lm(pm25 ~ temperature + humidity + wind, data = pm25)
anova_pm25.2 <- anova(pm25.lm2)
print(anova_pm25.2)
```

```
## Analysis of Variance Table
##
## Response: pm25
##             Df  Sum Sq Mean Sq  F value    Pr(>F)
## temperature  1  9014.4  9014.4  90.4498 5.713e-13 ***
## humidity     1 12739.7 12739.7 127.8296 1.266e-15 ***
## wind         1   622.6   622.6   6.2475   0.01563 *
## Residuals   52  5182.4    99.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comments: Every predictor in the new model is all significant. Therefore, we can use this new model instead of the full model.

e) Find the R-squared and comment on what it means in the context of this data set. 1/2

$$R^2 = 0.8127448$$

Comments: This high R-squared value shows that the independent variables (temperature, humidity, precipitation, and wind) contribute enormously to the variation of PM2.5 concentration.

Not quite right

f) The best multiple regression model that explains the data. 1/3

```r
summary(pm25.lm)$coefficients
```

```
##                   Estimate  Std. Error     t value     Pr(>|t|)
## (Intercept)   102.72258771 14.71952825   6.9786603 5.881953e-09
## temperature     1.62141831  0.18762464   8.6418198 1.463129e-11
## humidity       -1.27742262  0.11854373 -10.7759612 9.490343e-15
## precipitation  -0.01090918  0.02349567  -0.4643059 6.444046e-01
## wind           -0.58015926  0.23405331  -2.4787484 1.653279e-02
```

Comments: Variable precipitation has the largest P-Value, therefore, it explains the least variation when added to the model, so we should drop this variable from the model.

You answered this question in part g, you need to give more detail

g) Comment on the R-squared and adjusted R-squared in the full and final model you chose in part f. 3/3

Comments: The R-squared in the full model is higher than the reduced model ($0.8127 > 0.812$) because we can always increase R-squared by adding more predictors into the model. However, the adjusted R-squared is more accurate because it tries to balance the R-squared with the number of predictors by penalising for the number of parameters. As can be seen, the adjusted R-squared in the reduced model is higher than the full model ($0.8011 > 0.7981$), which means the reduced model is more reliable.

Question 2 19/25

a) For this study, is the design balanced or unbalanced? Explain why. 2/2

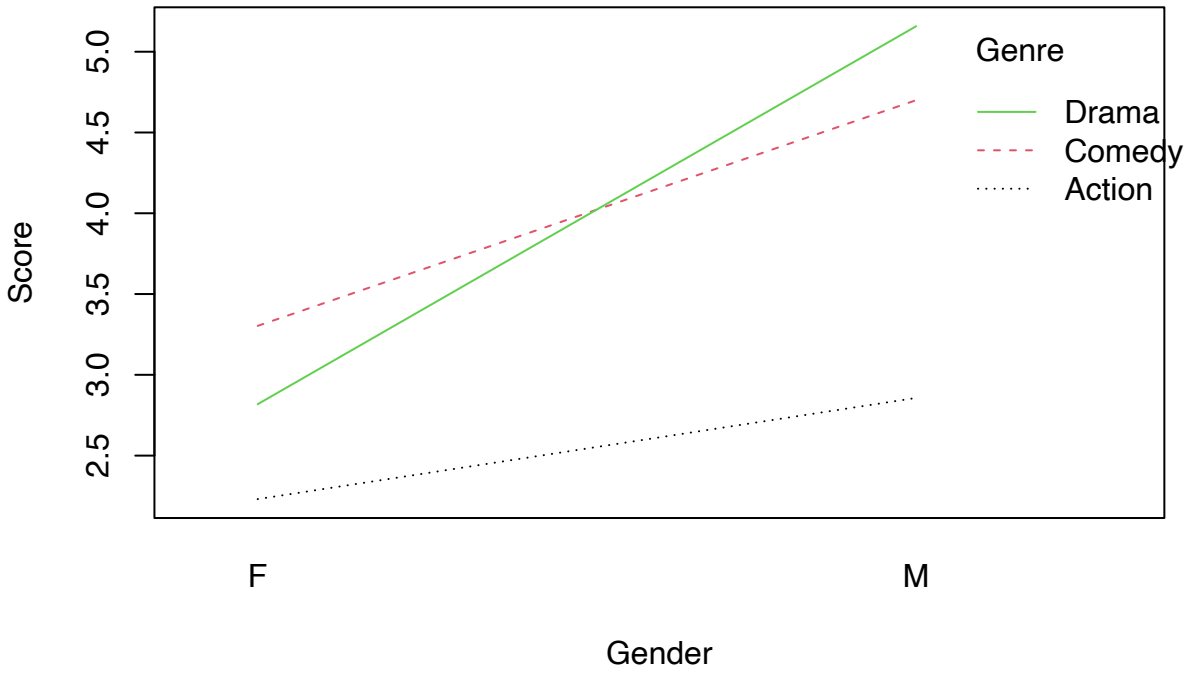
```
table(movie[, 1:2])
```

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

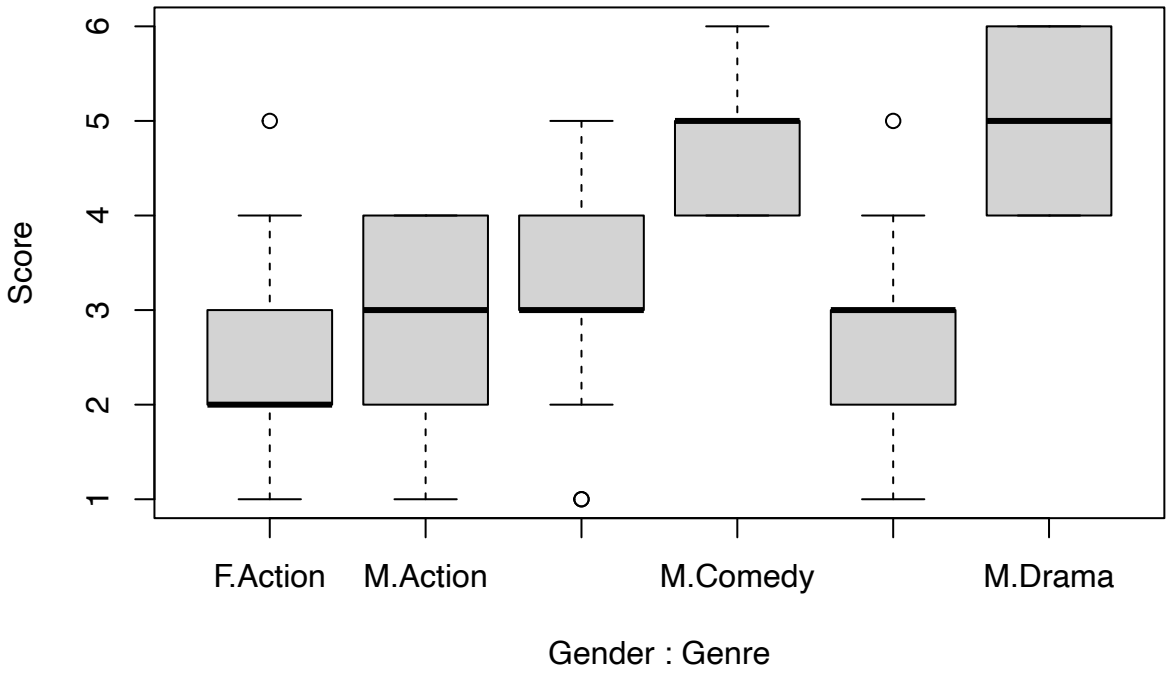
Comments: This is an unbalanced design, since the number of r replicates for each pair of factors are not the same.

b) Construct two different preliminary graphs that investigate different features of the data and comment. 6/8

```
with(movie, interaction.plot(Gender, Genre, Score, ylab = "Score", col =1:3))
```




```
boxplot(Score ~ Gender + Genre, data = movie)
```



Comments: The interaction plot indicates that there might be an interaction between the factors drama and comedy since they are not parallel. On the contrary, there is no interaction between action and comedy because they are parallel. Besides, the box plot shows that there is no visible constant variability between levels of each effect. Also, we can observe three outliers in the plot. Therefore, we should interpret the result with caution.

Need to review plots

4/4 **c) Write down the full mathematical model for this situation, defining all appropriate parameters.**

$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$

Where:
Y_{ijk} is the observed value of the dependent variable.
μ is the overall mean of the dependent variable.
α_i is the effect of the i-th level of factor A.
β_j is the effect of the j-th level of factor B.
γ_{ij} represents the interaction effect between factor A and factor B.
ε_{ijk} is the random unexplained variation.

Please give more detail in future

d) 7/9

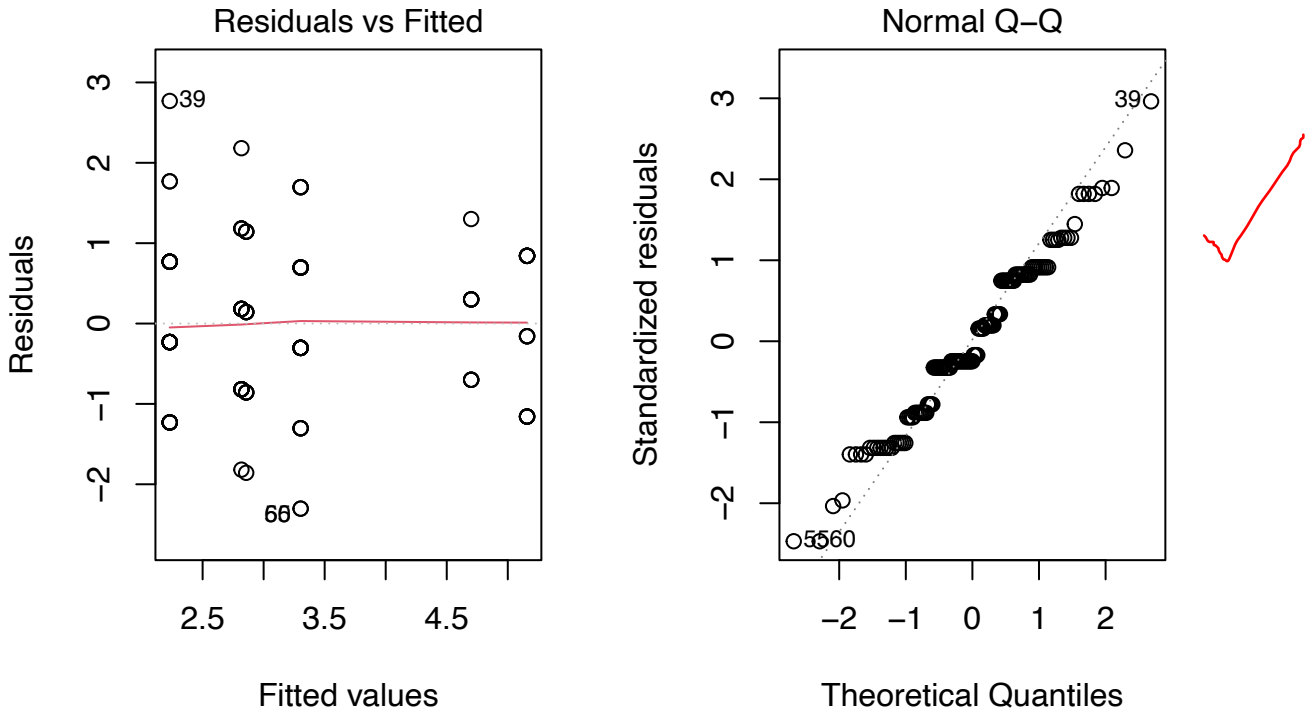
Hypotheses:

For factor Gender: $H_0 : \alpha_i = 0$ for all i; H_1 : at least one $\alpha_i \neq 0$

For factor Genre: $H_0 : \beta_j = 0$ for all j; H_1 : at least one $\beta_j \neq 0$

True but this should only be 1 null and alternative hypothesis

```
movie.aov = lm(Score ~ Gender* Genre, data = movie)
par(mfrow =c(1,2))
plot(movie.aov, which = 1:2)
```



Comments: The data point in the residuals vs fitted plot scatter evenly along the vertical and horizontal axis. Although there are some deviations, the residuals in the normal Q-Q plot still demonstrate a linear trend, which means that the residuals are close to normally distributed. Therefore, the Two-Way ANOVA assumptions are satisfied.

```
anova(movie.aov)
```

```
## Analysis of Variance Table
##
## Response: Score
##               Df  Sum Sq Mean Sq F value    Pr(>F)
## Gender         1  71.583  71.583 79.8038 3.277e-15 ***
## Genre          2  50.357  25.178 28.0698 7.152e-11 ***
## Gender:Genre   2  15.079   7.540  8.4054 0.0003677 ***
## Residuals    131 117.506   0.897
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comments: Because the P-Value of the interaction term (0.0003677) is lower than 0.01, we have enough evidence to conclude that there is significant interaction between factors Gender and Genre. Because the P-Value of factor Gender and factor Genre are lower than 0.01, we have sufficient evidence to reject the null hypothesis and conclude that factor Gender and factor Genre have significant effects on the response Score.

You should reason more about you conclusion, you have reached the final model.

e) Practical implications for the business that aims to maximise the brand recognition from the placement. Advice/Interpretation on the effect drama genre on the brand 0/2

Suggestions for Improvement: Share at least one recommendation on how the model solution and marking guide can be improved.Your in