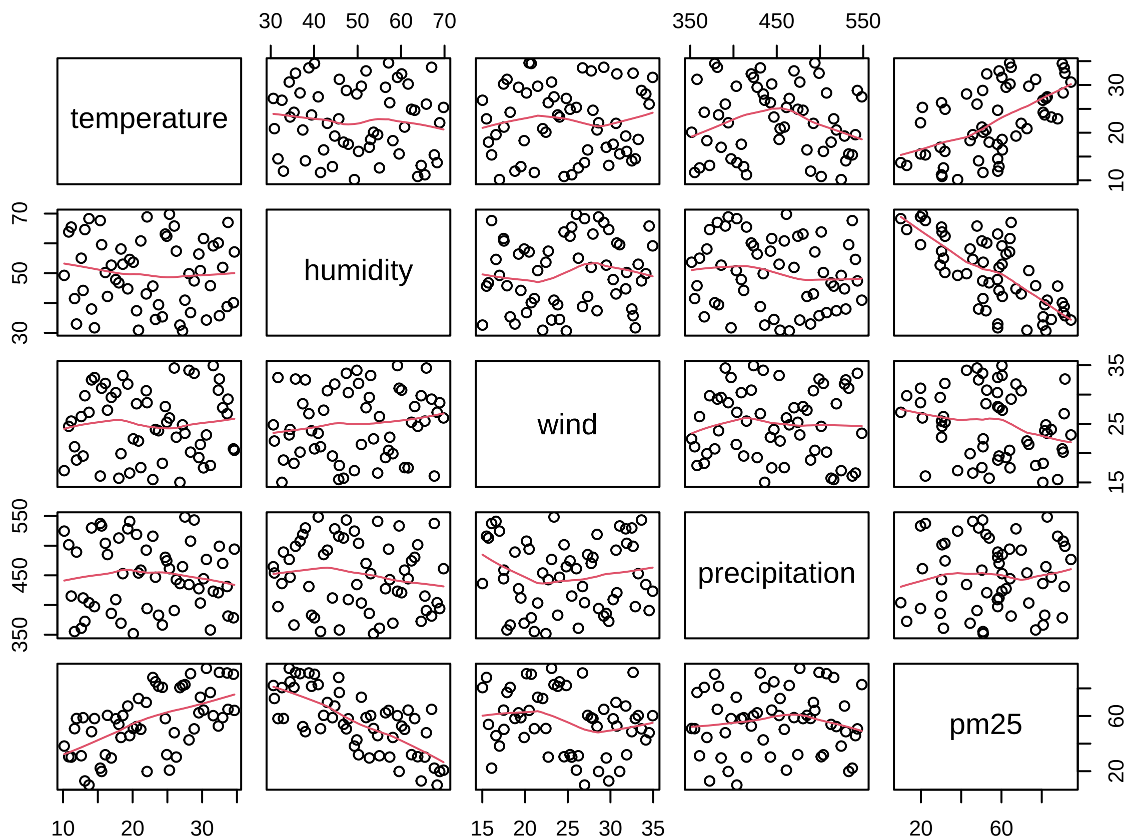
Question 1

a. Inputting the data and producing the scatterplot and correlation matrix:

```r
pm25 <- read.csv(here::here("assignment", "pm25.csv"))
pairs(pm25, panel = panel.smooth)
```




```r
cor(pm25)
```

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

- The response variable `pm25` has a moderate positive linear relationship with the predictor `temperature`; a moderately strong negative linear relationship with the predictor `humidity`; a weak negative linear relationship with the predictor `wind`. The response `pm25` has no obvious relationship with the predictor `precipitation`.

- There doesn't seem to be a relationship present between the predictors themselves.

  [**7 marks**] : 2 for producing the graph, 1 for producing the correlation matrix, 2 for comment on response v predictors, 2 for comment on among predictors

b. Fit the full model:

```
M1 <- lm(pm25 ~ ., data = pm25)
summary(M1)
```

```
#
# Call:
# lm(formula = pm25 ~ ., 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 ***
# wind           -0.58016    0.23405  -2.479   0.0165 *
# precipitation  -0.01091    0.02350  -0.464   0.6444
# ---
# 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
```

```
summary.M1 <- summary(M1)
se <- sqrt(diag(summary.M1$cov.unscaled * summary.M1$sigma^2))[3]
```

The required CI is

$$
\begin{aligned}
&\hat{\beta}_{\text{humidity}} \pm t_{n-p,1-\alpha/2} s.e.(\hat{\beta}_{\text{humidity}}) \\
=&\hat{\beta}_{\text{humidity}} \pm t_{51,0.975} s.e.(\hat{\beta}_{\text{humidity}}) \\
=&-1.277 \pm 2.0075838 \times 0.1185437 \\
=&(-1.5149865, -1.0390135).
\end{aligned}
$$

That is, we are 95% confident that for every percentage increase in relative humidity, the $PM_{2.5}$ concentration will decrease between 1.0390135 and 1.5149865 milligram per cubic meter ($\mu g/m^3$) on average.

[**6 marks**] : 1 for fitting the full model; 1 for correct $\hat{\beta}$; 1 for correct s.e.; 1 for correct quantile; 1 for correct calculation; 1 for comment

c. - Theoretical Model is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \epsilon_i, \quad i = 1, 2, \ldots n$$

- Y is the response variable `pm25`;
- X_{ij} are the predictors variables for the i-th observation:
  - X_{i1} = annual mean `temperature` of test locations
  - X_{i2} = annual mean `humidity` of test locations
  - X_{i3} = annual mean `wind` speed of test locations
  - X_{i4} = annual mean `precipitation` of test locations
- $\epsilon \sim N(0, \sigma^2)$ denotes the random variation with constant variance;

[**5 marks**] : For defining the model and its parts. 1 for the model equation, 1 mark for defining the response, 2 marks for defining the predictors (so 0.5 each), 1 mark for definition of the random variation.

Conducting the F-test we have,

- Hypotheses: $H_0 : \beta_1 = \ldots = \beta_4 = 0$ vs H_1 : not all $\beta_i = 0$; $i = 1, 2, \ldots, 4$.
- Standard R output ANOVA table

```
anova(M1)
```

```
# 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 ***
# wind           1   622.6   622.6   6.1533   0.01646 *
# precipitation  1    21.8    21.8   0.2156   0.64440
# Residuals     51  5160.6   101.2
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Can show the reduced Overall ANOVA table as

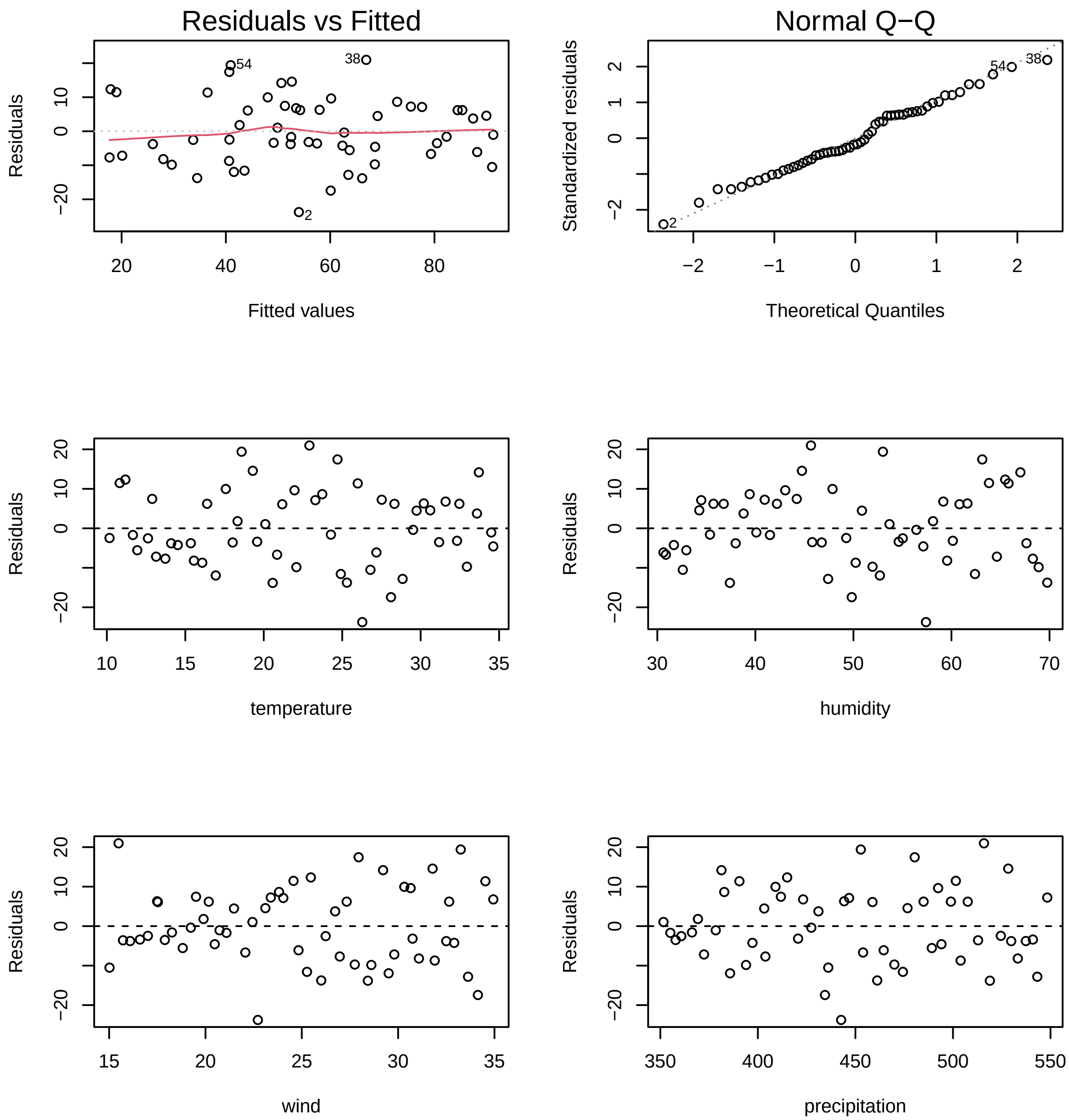
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Regression	4	22398.593	5599.6483	55.3388806420261	0
Residuals	51	5160.604	101.1883		

- Note the Regression SS $= 9014.4 + 12739.7 + 622.6 + 21.8 = 22398.59$
- Therefore the Mean Square Reg = Reg SS/Reg df $= 22398.59/4 = 5599.64832$
- Test statistic: $F_{obs} = MS_{Reg}/MS_{Res} = 5599.64832/101.188319 = 55.3388806$;
- The null distribution for the test statistics is $F_{4,51}$.
- P-value: $P(F_{4,51} \geq 55.3388806) = 0 = 6.0817191 \times 10^{-18} < 0.05$;
- As the P-value is small,
  - (Statistical) There is enough evidence to reject H_0.
  - (Contextual) That is, there is a significant linear relationship between `pm25` and at least one of the 4 predictor variables.

[**9 marks**] : For conducting the test above. 1 for the hypotheses; 5 for the ANOVA table (i.e. 1 per column). 1 for stating the null distribution of the test statistics explicitly; 2 for the conclusion.

d. For the diagnostics:

```r
par(mfrow = c(3, 2))
plot(M1, which = 1:2)
plot(resid(M1) ~ temperature, data = pm25, xlab = "temperature", ylab = "Residuals")
abline(h = 0, lty = 2)
plot(resid(M1) ~ humidity, data = pm25, xlab = "humidity", ylab = "Residuals")
abline(h = 0, lty = 2)
plot(resid(M1) ~ wind, data = pm25, xlab = "wind", ylab = "Residuals")
abline(h = 0, lty = 2)
plot(resid(M1) ~ precipitation, data = pm25, xlab = "precipitation", ylab = "Residuals")
abline(h = 0, lty = 2)
```



- The quantile plot of residuals look approximately linear, suggesting the normality assumption for residuals is appropriate;

- There is no obvious pattern in any of the residual plots so it appears the linearity and constant variance assumptions of the multiple linear model are justified.

[**10 marks**] : 1 for each of the six plots ('abline' again is optional), 2 for commenting on the qq-plot, 2 for commenting on the residual plots

e. Here $R^2 = 0.813 = 81.3\%$, which is a goodness of fit metric. It means 81.3% of the variation in pm25 is explained by the full linear regression model.

[**2 marks**] : 1 For correct value and 1 for contextual statement

f. Starting with all the predictors

```
summary(M1)
```

```
#
# Call:
# lm(formula = pm25 ~ ., 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 ***
# wind           -0.58016    0.23405  -2.479   0.0165 *
# precipitation  -0.01091    0.02350  -0.464   0.6444
# ---
# 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
```

precipitation has the highest P-value so we shall remove it first.

```
M2 <- update(M1, . ~ . - precipitation)
summary(M2)
```

```
#
# Call:
# lm(formula = pm25 ~ temperature + humidity + wind, data = pm25)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -23.7588  -6.4368  -0.5659   6.4006  20.2813
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)  97.3234     8.9561  10.867 5.45e-15 ***
# temperature   1.6267     0.1859   8.753 8.39e-12 ***
# humidity     -1.2698     0.1165 -10.899 4.89e-15 ***
```

```
# wind         -0.5806     0.2323  -2.500   0.0156 *
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 9.983 on 52 degrees of freedom
# Multiple R-squared:  0.812,   Adjusted R-squared:  0.8011
# F-statistic: 74.84 on 3 and 52 DF,  p-value: < 2.2e-16
```

At this point, all remaining predictors are significant and should be kept in the model. The final (fitted) model equation is

$\hat{Y} = 97.323 + 1.627X_1 - 1.27X_2 - 0.581X_3$ or $\widehat{\texttt{pm25}} = 97.323 + 1.627\texttt{temperature} - 1.27\texttt{humidity} - 0.581\texttt{wind}$.

[**3 marks**] : 2 in total for any correct procedure; 1 for the fitted model equation

f. The R^2 goodness of fit metric always decreases/increases when a predictor is removed/added from/into the model. The adjusted R^2 has a penalty for the number of predictors in the model. So it will sometimes increase when a predictor is removed. In this case, from the full to final model, the R^2 decreases from 81.3%to 81.2% but the adjusted R^2 increases from 79.8% to 80.1%. This indicates the final model is a better parsimonious model for the data.

   [**3 marks**] : 2 marks for correctly comparing values between models. 1 mark for explaining there's a penalty on parameters in adjusted R^2.

Question 1 Total marks: 45

Question 2

a. A study is balanced if there are equal number of replicates across all the levels factors in the study. Here we check the number of replicates with,

```r
movie <- read.csv(here::here("assignment", "movie.csv"),
  header = TRUE,
  stringsAsFactors = TRUE
)
table(movie[, c("Gender", "Genre")])
```

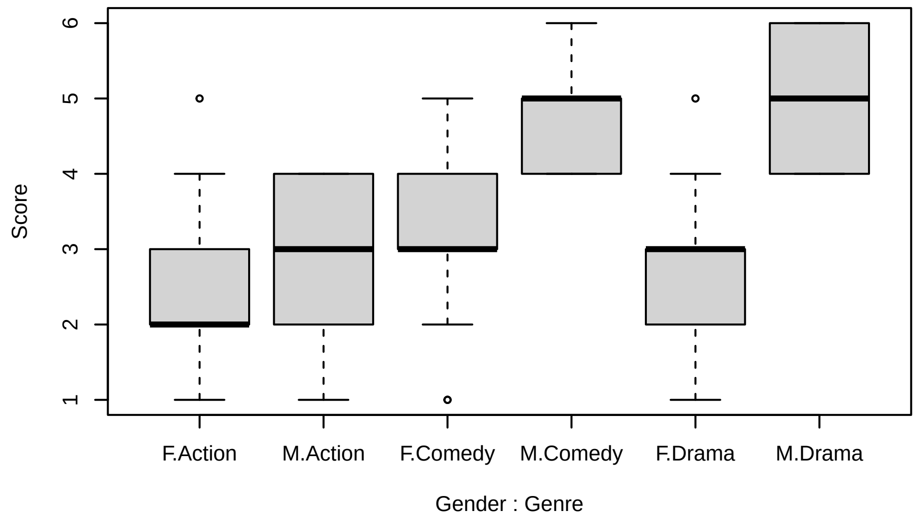
```
#       Genre
# Gender Action Comedy Drama
#      F     39     33    22
#      M     14     10    19
```

   From the above we can see that the design is unbalanced with an unequal number of replicates for each combination of levels of the two factors.

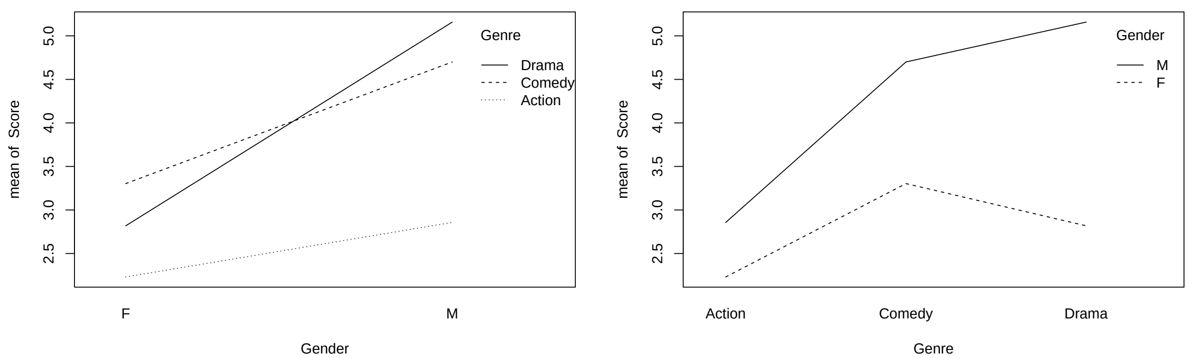
   [**2 marks**] : 1 mark for computing replicates; 1 mark for explaining balanced design has equal replicates

b. Constructing the preliminary plots

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




```r
par(mfrow = c(1, 2))
with(movie, interaction.plot(Gender, Genre, Score))
with(movie, interaction.plot(Genre, Gender, Score))
```



- From both interaction plots we can see non-parallel lines for the means of each group at different levels of the independent variables, this indicates a significant interaction effect between the two independent variables.
- From the boxplot, we can see that the assumption of equal variance among levels seems approximately valid due to the similar box sizes. Optionally we can also compute the standard deviation for each group:

```
#   Gender  Genre         s
# 1      F Action 0.9308044
# 2      F Comedy 1.0453722
# 3      F  Drama 0.9579921
# 4      M Action 0.9492623
# 5      M Comedy 0.6749486
# 6      M  Drama 0.8983416
```

The standard deviations are quite similar.

[**8 marks**] : 2 marks for a decent boxplot; 2 marks for commenting on the boxplot; 2 marks for at least one of the two interaction plots above; 1 mark for noticing close to non-parallel lines/non-constant slopes; 1 mark for concluding about interaction

c. The full Two-Way ANOVA model with interaction is:

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

where the parameters are :

- Y_{ijk}: the brand recall score response;
- α_i: the `Gender` effect, there are two levels - Male and Female
- β_j: the `Genre` effect, there are three levels - Action, Comedy and Drama;
- γ_{ij}: interaction effect between `Gender` and `Genre`.
- $\epsilon_{ijk} \sim N(0, \sigma^2)$ is the unexplained variation.

[**4 marks**] : 1 mark for writing the full model correctly; 0.5 mark each for defining Y, α, β, γ; and 1 mark for defining ϵ

d. We wish to first test

$$H_0 : \gamma_{ij} = 0 \text{ for all } i, j \quad \text{against} \quad H_1 : \text{at least one } \gamma_{ij} \neq 0$$

Fitting this interaction model

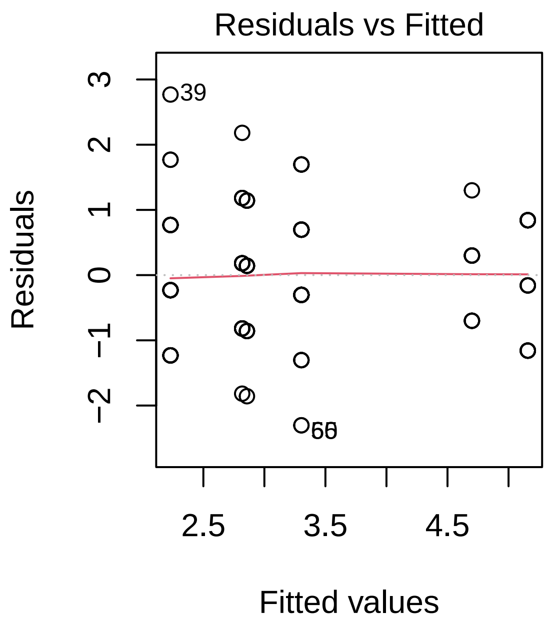
```
movie.int <- lm(Score ~ Genre * Gender, data = movie)
anova(movie.int)
```

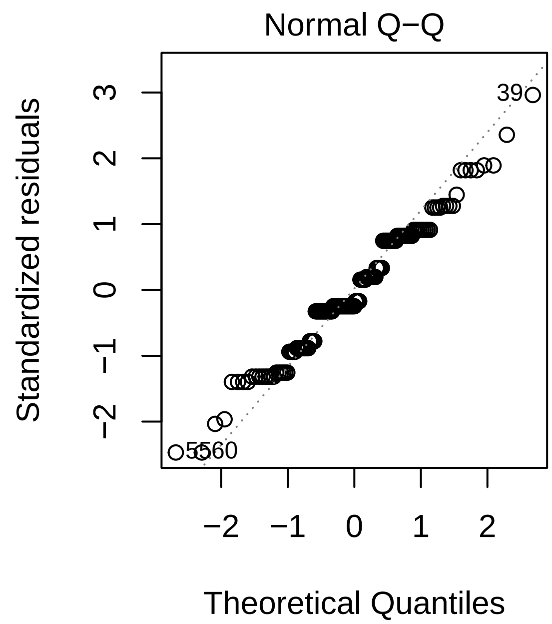
```
# Analysis of Variance Table
#
# Response: Score
#               Df  Sum Sq Mean Sq F value    Pr(>F)
# Genre          2  62.190  31.095 34.6658 8.254e-13 ***
# Gender         1  59.750  59.750 66.6117 2.388e-13 ***
# Genre:Gender   2  15.079   7.540  8.4054 0.0003677 ***
# Residuals    131 117.506   0.897
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- We can see that the interaction terms are significant since the F-test of the interaction term has a P-value of $3.677164 \times 10^{-4} < 0.05$, they can't be removed from the model. This also means we reached our final model.

We should validate the interaction model with the diagnostic plots.

```
par(mfrow = c(1, 2))
plot(movie.int, which = 1:2)
```





The residuals are close to linear in the QQ-plot, and so the normal assumption should be valid. The residual plot seems to show equal spread around the fitted values and so the constant variance assumption is also appropriate.

[**9 marks**] : 1 mark for fitting the model; 1 for the ANOVA table; 1 mark for the correct hypotheses, 1 mark for noticing the interaction terms are significant; 1 mark for stating the interaction terms can't be removed and reached the final model; 2 marks for producing the plots; 2 marks for comments

e. Overall, the effect of the *gender* of the audience on brand recall *score* depends on the movie *genre*. Male recall more brands when they watch drama, and female recall more brands when they watch comedy. It also shows that difference in brand recall *score* is reinforced/amplified when the *genre* is drama for both males and females.

We can't interpret the effect of drama alone due to the significant interaction effect between movie *genre* and *gender* of the audience.

[**2 marks**] : 1 mark for comments on practical implications to the business; 1 mark for mentioning we can't interpret the main effect due to significant interaction effect

Question 2 Total marks: 25