Applied_Statistics

odi

2023-05-16

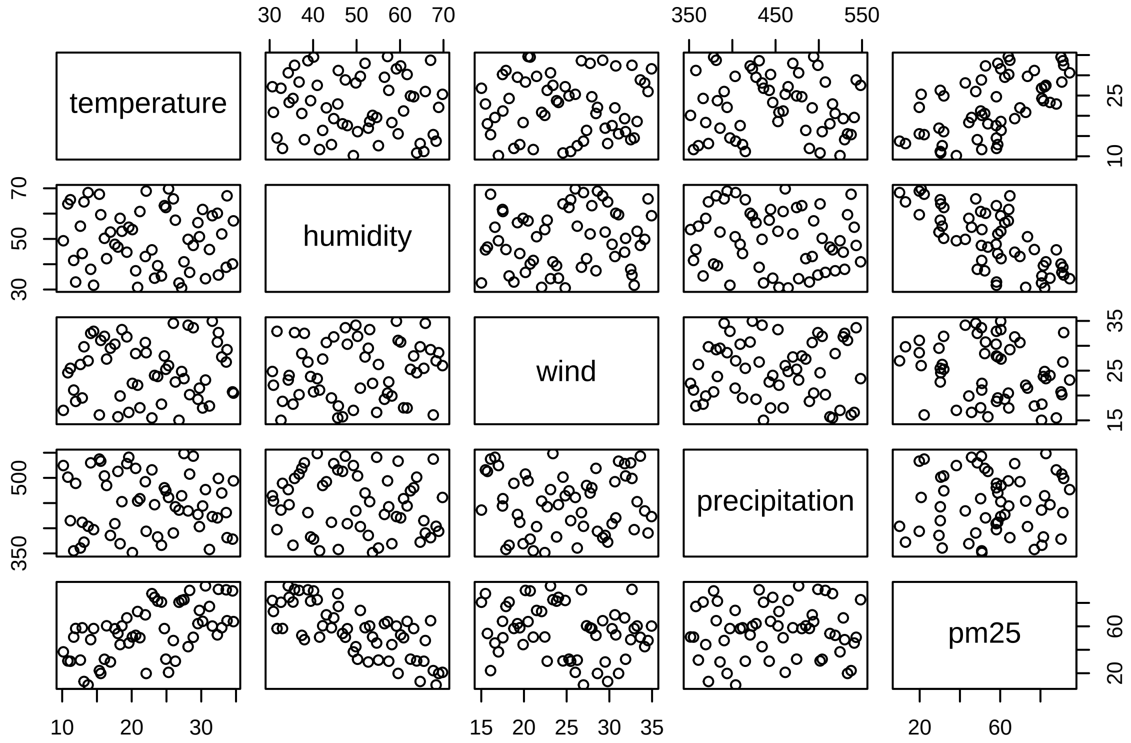
Part 1

The analysis of PM2.5 concentrations in relation to meteorological factors at 56 test locations involves several steps. First, scatter plots and a correlation matrix reveal relationships between PM2.5 and predictors (temperature, humidity, wind speed, precipitation), as well as among predictors. A multiple regression model is then fit to quantify these impacts, and a 95% confidence interval is used to estimate the effect of humidity on PM2.5. An F-test assesses the overall significance of the model, with a significant p-value indicating that the predictors collectively impact PM2.5 levels. Model validation includes residual analysis and checking for multicollinearity and goodness-of-fit metrics. Using model selection procedures, the best multiple regression model is identified. Comparing R^2 and adjusted R^2 between the full and final models helps to understand the efficiency and explanatory power of the chosen model, leading to a comprehensive approach for predicting PM2.5 concentrations based on meteorological data.

#loaded libraries required to produce the outputs
options(repos = “https://cloud.r-project.org”)
install.packages(“rlang”)
install.packages(“ggplot2”) library(ggplot2)
library(corrplot)

```r
#read the dataset
data <- read.csv("data/pm25.csv", header = TRUE)
```

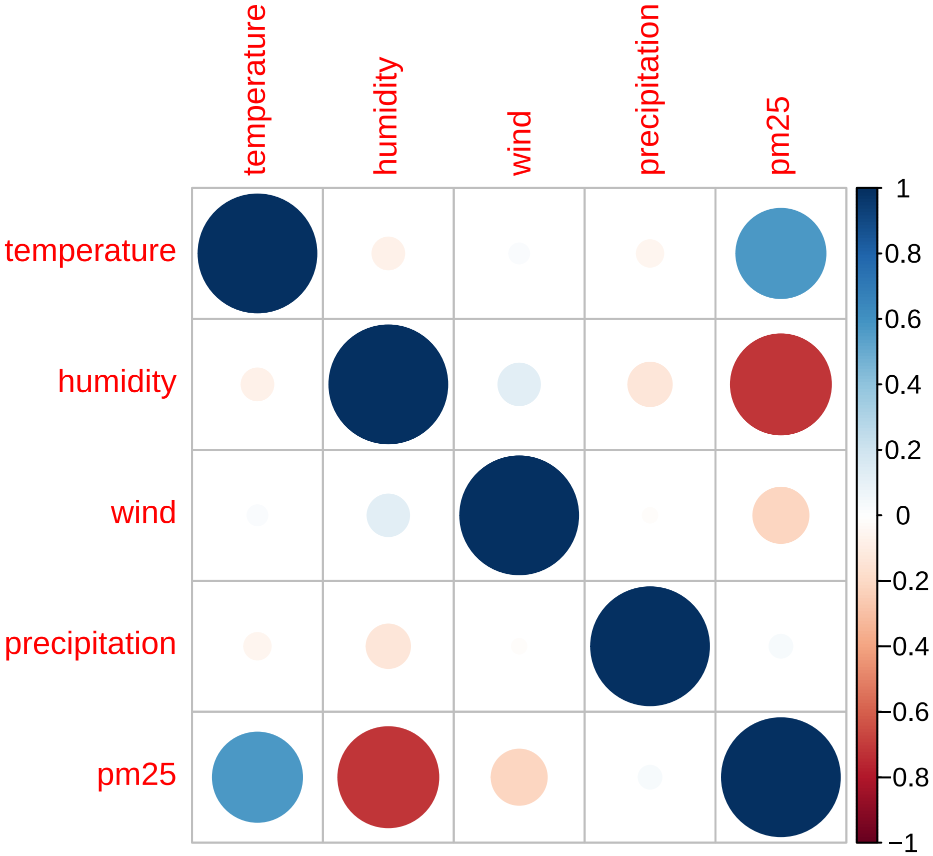
```r
#Displayed Scatter plot matrix of all variables
plot(data)
```



```r
#Displayed Correlation matrix of all variables
cor_matrix <- cor(data[, c("temperature", "humidity", "wind", "precipitation", "pm25")])
cor_matrix
```

```
##               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
```

```r
#Plotted the correlation matrix
corrplot::corrplot(cor_matrix, method = "circle")
```



Upon close examination of the scatter plot, correlation matrix and the correlation plot above,there appears to be a **positive** linear relationship between **temperature** and **pm25**, meaning that as temperature increases, so does pm25. On the other hand, there appears to be a **negative** linear relationship between **humidity** and **pm25**, meaning that as humidity increases, **pm25** decreases, and vice-versa. The last relationship worth noting is how **wind** has a **low negative** correlation with pm25. Every other relationship between any two variables are not worth mentioning since these are all extremely weak relationships.

Multiple Linear Regression Model

In order for all the predictors to be accounted for, I will be using the **multiple linear regression model**.

```
#fitted the model
data_model <- lm(pm25 ~ temperature + humidity + wind + precipitation, data = data)

#got summary of the model
summary(data_model)
```

```
##
## Call:
## lm(formula = pm25 ~ temperature + humidity + wind + precipitation,
##     data = data)
##
## 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
```

Humidity Impact

```r
#Got the coefficient and confidence interval for humidity
coef <- coef(data_model)["humidity"]
ci <- confint(data_model)["humidity",]

# Displayed the coefficient and confidence interval
cat("Coefficient for humidity:", coef)
```

```
## Coefficient for humidity: -1.277423
```

```r
cat("95% confidence interval:", ci)
```

```
## 95% confidence interval: -1.515409 -1.039436
```

This means that, holding all other predictors constant, for each extra percentage of relative humidity, we would expect PM25 concentration to decrease by -1.277423. The 95% confidence interval ranges from -1.515409 to -1.039436, which means that we can be 95% confident that the true effect of humidity on PM25 concentration is within this range, hence, supporting the fact that -1.277423 is within this range and as a result, we can say that we are 95% confident that the coefficient value of humidity is within the range.

The multiple linear regression model can be written as:

$$pm25 = \beta 0 + \beta 1 temperature + \beta 2 humidity + \beta 3 wind + \beta 4 precipitation + \epsilon$$

where $\beta 0$ is the intercept, $\beta 1$ to $\beta 4$ are the coefficients for each predictor variable, and ϵ is the error term. The coefficients represent the change in pm25 for a unit change in each of the predictor variables, holding all other predictors constant.

- Write down the Hypotheses for the Overall ANOVA test of multiple regression.

The null hypothesis for the overall ANOVA test of multiple regression is that for every change in any of the predictors, the change in the response variable or y, is equal to 0, meaning there is no change, hinting that there is no relationship between the response variable and any of the predictor variables:

$$H0 : \beta 1 = \beta 2 = \beta 3 = \beta 4 = 0$$

The alternative hypothesis is that there is at least one coefficient of a predictor variable that is not equal to 0, hinting that there is a relationship between the response variable and the predictor variable/s with coefficients not equal to 0:

$$Ha : \beta j \neq 0$$

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

```
#Got ANOVA table for the model

aov_table <- anova(data_model)
aov_table
```

```
## 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
```

Using the output from the table above, I computed for the following:

Full Model Regression

DF : 4
SS : 9014.39411 + 12739.74381 + 622.64119 + 21.81417 = 22398.59
MS: (SS/DF) 22398.59 / 4 = 5599.648

Error

DF: 51
SS: 5160.60428
MS: (SS/DF) 5160.60428 / 51 = 101.1883

F-Value

F-value:(MS of Full Model Regression / MS of Error) 5599.648 / 101.1883 = 55.33889

Total

DF: (Full Model DF + Error DF) 4 + 51 = 55
SS: (FUll Model SS + Error SS) 22398.59 + 5160.60428 = 27559.19

Using the output from the computations above, I will be producing the ANOVA table for overall multiple regression model where one combined regression SS is displayed.

```
#Created data frame for the new ANOVA table

anova_table <- data.frame(Source = c("Model", "Error", "Total"),Df = c(4, 51,55), Sum_Sq = c(22398.59, 5

#Displayed the new ANOVA table
print(anova_table, row.names = FALSE)
```

```
##  Source Df    Sum_Sq   Mean_Sq  F_value
##   Model  4 22398.590 5599.6480 55.33889
##   Error 51  5160.604  101.1883       NA
##   Total 55 27559.190        NA       NA
```

- Compute the F statistic for this test.

The F statistic for this test is 55.33889, with 4 numerator degrees of freedom and 51 denominator degrees of freedom.

- State the Null distribution for the test statistic.

The null distribution for the test statistic is an F-distribution with (4, 51) degrees of freedom.

- Compute the P-Value

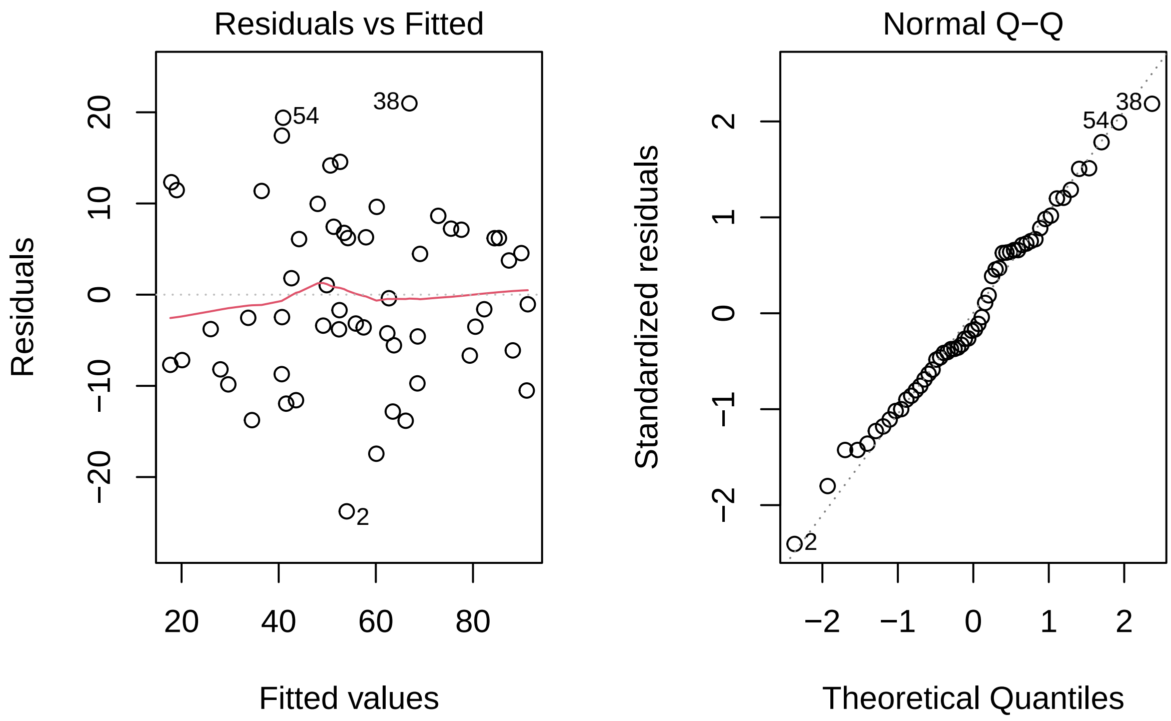
```
#Calculated the P-value

p_value <- 1 - pf(55.33889, 4, 51)

#Displayed the P-value

print(p_value)
```

```
## [1] 0
```

- State your conclusion (both statistical conclusion and contextual conclusion).

The p-value is a very small number that was rounded off to 0, which means that it is significantly less than 0.05, indicating strong evidence against the null hypothesis. Therefore, we can reject the null hypothesis and conclude that there is a significant relationship between the response variable PM25 and at least one of the predictors (temperature, humidity, wind, and precipitation). Given this result, it is evident that the model is reliable in predicting PM25 concentration based on the mentioned predictors.

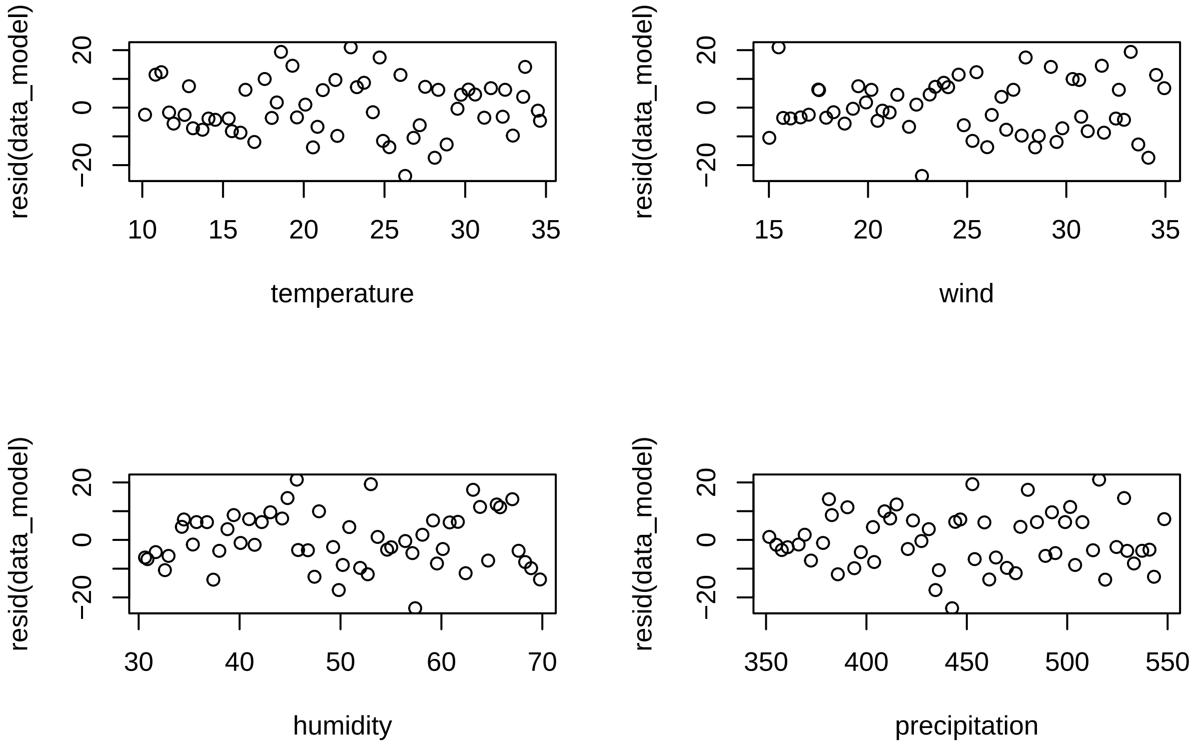
```r
#created scatterplot of residuals and fitted
par(mfrow = c(1, 2))
plot(data_model, which = 1:2)
```



Based on the residuals vs fitted plot, residuals are randomly scattered with no discernible pattern. This suggests that the model may be appropriate.

The Normal Q-Q plot of residuals, has slight curvature but close to depicting a straight line, implying residuals/errors are close to a normal distribution, hence, normality can be assumed and that the model may be deemed appropriate.

```r
#Created Residuals VS Predictor Plots
par(mfrow = c(2, 2))
plot(resid(data_model) ~ temperature + wind + humidity + precipitation, data = data)
```



Residuals vs predictor plots show no obvious pattern and points appear to be randomly spread which also supports that model is appropriate.

Given these findings, it can be said that full regression model passed the tests, and it can be concluded that it is appropriate to explain the PM2.5 concentration at various test locations.

```r
#Got the summary which contains the R2 of the model

summary(data_model)
```

```
##
## Call:
## lm(formula = pm25 ~ temperature + humidity + wind + precipitation,
##     data = data)
##
## 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
```

The R2 value for the model is 0.8127. Since an R2 value of 1 indicates a perfect fit, while an R-squared value of 0 indicates that the model does not explain any of the variation in the response variable; given the value of .8127 which is close to 1, indicates that the model is a good fit for the data. This means that the model explains about 81.27% of the variation in PM25 concentration at the various test locations.

For this part, I will first be selecting the best model using R alone. Then, for comparison, I will be conducting the procedures discussed in the course.

Stepwise selection by R

```
#applied step function in R
step_model <- step(data_model, direction = "both")
```

```
## Start:  AIC=263.31
## pm25 ~ temperature + humidity + wind + precipitation
##
##                 Df Sum of Sq     RSS    AIC
## - precipitation  1      21.8  5182.4 261.55
## <none>                        5160.6 263.31
## - wind           1     621.7  5782.3 267.68
## - temperature    1    7556.8 12717.5 311.82
## - humidity       1   11750.1 16910.7 327.78
##
## Step:  AIC=261.55
## pm25 ~ temperature + humidity + wind
##
##                 Df Sum of Sq     RSS    AIC
## <none>                        5182.4 261.55
## + precipitation  1      21.8  5160.6 263.31
## - wind           1     622.6  5805.1 265.90
## - temperature    1    7635.2 12817.6 310.26
## - humidity       1   11838.8 17021.2 326.14
```

```
summary(step_model)
```

```
##
## Call:
## lm(formula = pm25 ~ temperature + humidity + wind, data = data)
##
## 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
```

$$InitialModel : pm25 \sim temperature + humidity + wind + precipitation$$

$$FinalModel : pm25 \sim temperature + humidity + wind$$

The best model selected by the stepwise regression procedure includes only temperature ,humidity and wind as predictor variables.

The final fitted regression model is:

$$pm25 = 97.3234 + 1.6267 * temperature - 1.2698 * humidity - 0.5806 * wind$$

This model suggests that for every one degree Celsius increase in temperature, the PM2.5 concentration increases by 1.6267. On the other hand, for every increase in humidity, the PM2.5 concentration decreases by -1.2698. Similarly, for every change in wind, the PM2.5 concentration decreases by -0.5806. The model has an adjusted R-squared of 0.8011, indicating that it explains about 80% of the variation in the PM2.5 concentrations.

Below, I will be checking that if I indicated the direction to backward elimination, the result will be the same.

```r
#applied step with "backward" direction
backward_model <- step(data_model, direction = "backward")
```

```
## Start:  AIC=263.31
## pm25 ~ temperature + humidity + wind + precipitation
## 
##                 Df Sum of Sq     RSS    AIC
## - precipitation  1      21.8  5182.4 261.55
## <none>                        5160.6 263.31
## - wind           1     621.7  5782.3 267.68
## - temperature    1    7556.8 12717.5 311.82
## - humidity       1   11750.1 16910.7 327.78
## 
## Step:  AIC=261.55
## pm25 ~ temperature + humidity + wind
## 
##               Df Sum of Sq     RSS    AIC
## <none>                      5182.4 261.55
## - wind         1     622.6  5805.1 265.90
## - temperature  1    7635.2 12817.6 310.26
## - humidity     1   11838.8 17021.2 326.14
```

```r
summary(backward_model)
```

```
##
## Call:
## lm(formula = pm25 ~ temperature + humidity + wind, data = data)
##
## 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
```

From the output, it can be seen that the result is the same.

Dropping of Variable that Explains Least Variation (Largest P-value)

For this part, I will be performing the procedures below to compare if the resulting model is the same with the model selected above.

Step 1: Regress with all the predictor variables in the model
Step 2.1: Drop the variable with the largest p-value in the t-test
Step 2.2: Regress with the reduced model
Step 3: Run Steps 2.1-2.2 iteratively until all variables are significant

```r
summary(data_model)$coefficients
```

1st Iteration

```
##                    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
## wind           -0.58015926  0.23405331  -2.4787484 1.653279e-02
## precipitation  -0.01090918  0.02349567  -0.4643059 6.444046e-01
```

From the result, precipitation has the largest p-value with an insignificant value of 0.6444046, which is above 0.05. As a result, precipitation will be removed for this iteration.

```r
summary(lm(pm25 ~ temperature + humidity + wind, data = data))$coefficients
```

2nd Iteration

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 97.3233697  8.9561211  10.866688 5.449901e-15
## temperature  1.6267473  0.1858554   8.752758 8.392977e-12
## humidity    -1.2697664  0.1165025 -10.899052 4.890414e-15
## wind        -0.5805842  0.2322795  -2.499507 1.563113e-02
```

From the result, all variables or predictors remaining have p-values below 0.05, and as a result, no variable/predictor will further be removed as the remaining predictors are deemed significant to the model. The resulting model ended up with the same predictors above: temperature ,humidity and wind, and the same set of coefficient values; hence, the final fitted regression model is still as follows:

$$pm25 = 97.3234 + 1.6267 * temperature - 1.2698 * humidity - 0.5806 * wind$$

g. [3 marks] Comment on the R2 and adjusted R2 in the full and final model you chose in part f. In particular explain why those goodness of fitness measures change but not in the same way.

```r
#Got the summary of the final model

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

```
##
## Call:
## lm(formula = pm25 ~ temperature + humidity + wind, data = data)
##
## 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
```

Prior the removal of precipitation, the model, which includes all predictors, has an R-squared value of 0.8127. In the final model we obtained after the removal of precipitation, the adjusted R-squared value is 0.8011. The R-squared value for the model including all variables is higher than the adjusted R-squared value of the final model. This is an indication that the full model is able to explain more the response variable, PM2.5 as compared to the final model. However, the adjusted R-squared value takes into account the number of predictor variables in the model. Even if the final model has a lower R-squared value than the full model, it is important to note that it has a higher adjusted R-squared value. This suggests that including precipitation in the model did not significantly improve the model.

PART 2

A study was conducted to assess the impact of product placement in movies on brand recognition by recording the number of brands correctly identified by individuals who watched different genres of movies. The dataset included variables for gender, movie genre, and brand recall score. To determine if the study design was balanced, the number of observations across all gender and genre categories was checked. Two preliminary graphs—a bar plot showing the average brand recall score by gender and a box plot displaying the distribution of brand recall scores across different genres—were created to identify trends and differences. Hypothesis tests were conducted to analyze these effects, with null hypotheses stating that neither gender nor genre affects brand recall scores, and alternative hypotheses suggesting they do. Assumptions of normality and equal variances were checked. The analysis indicated that both gender and genre might influence brand recall. Practical implications for the business include focusing on movie genres that maximize brand recall.

```
#loaded the dataset

data_movie <- read.csv("data/movie.csv", header = TRUE)
```

```
#Displayed the table of the count of each observation belonging to each factor
table(data_movie[, c("Gender", "Genre")])
```

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

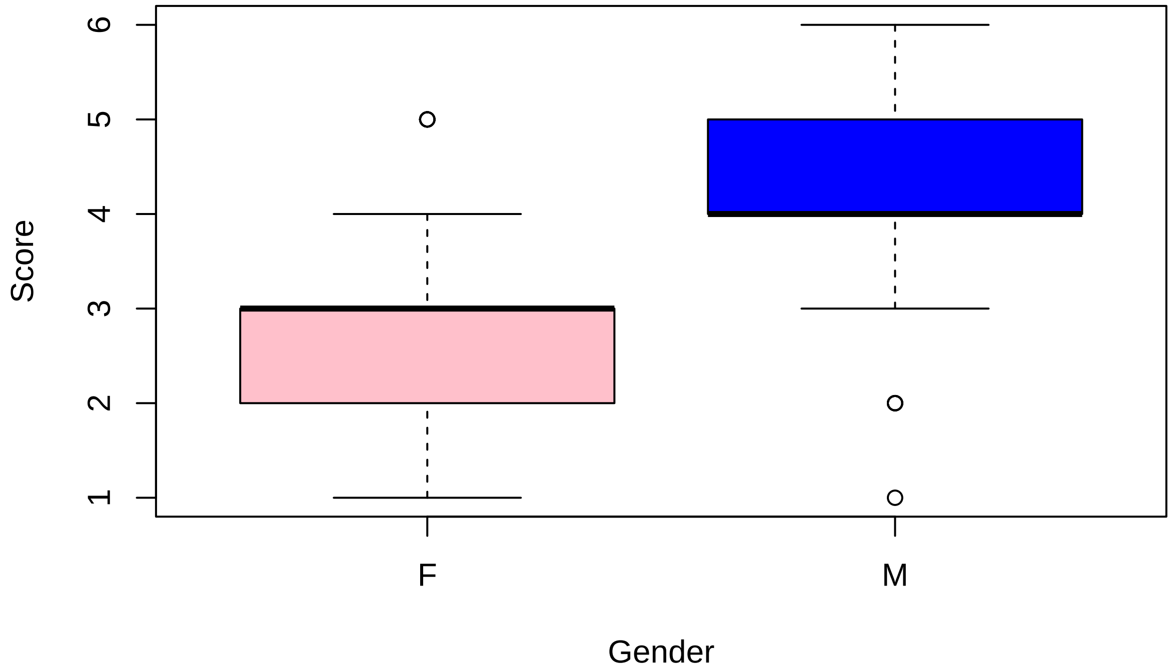
From the table above, it is evident that the design is unbalanced beacause a balanced study is when each treatment combination has the same number of observations; and in this case, we have an unequal number of observations between males and females; and an unequal number of observations among the different movie genres.With this said, the design is unbalanced because the different treatment combinations have different sample sizes for different levels of each factor combination.

Comparative Box Plot of Gender, and Genre by Score

First, I will be constructing a box plot of the scores against the factors Gender and Genre to check if there are any differences in terms of scores among these factors.

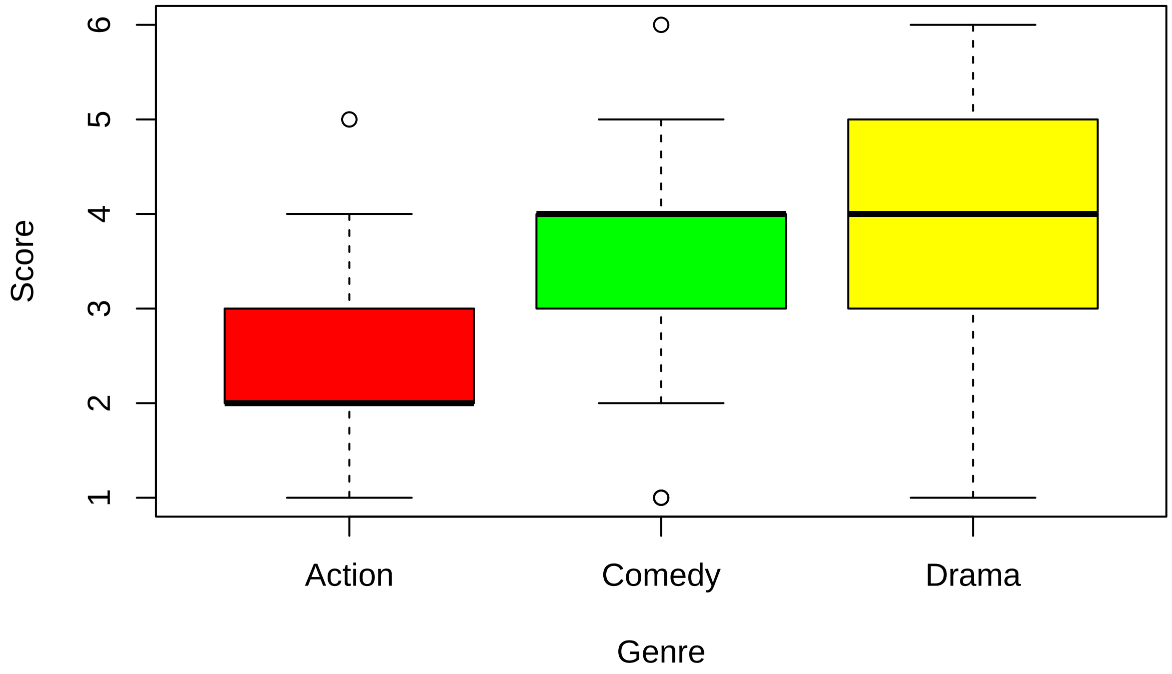
Before I create the box plot that will help us analyze the interaction between the Gender and Genre factor, to have a clearer view and have a better understanding of each variable individually, I will first make separate box plots where each factor is placed against the scores.

```
#created box plot for Gender against Scores

boxplot(Score ~ Gender, data = data_movie, col = c("pink","blue"))
```



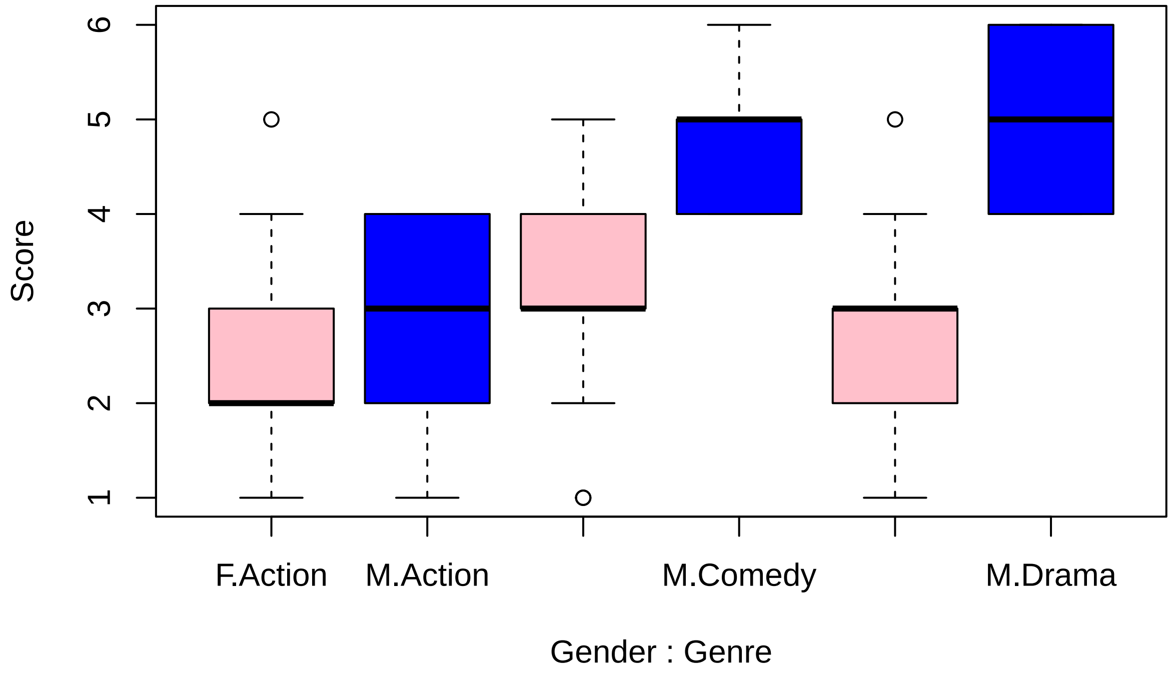
From the plot, we can see that there are in fact differences between Males and Females in terms of scores. It can be seen that males tend to recall more brands than females. Moreover, it is also important to note that though this is the case, it can be seen from the box plot that males and females tend to be both normally distributed in terms of scores, and that their variations are close to one another despite having different number of sample sizes.This could be an indication that gender is an important factor in brand recognition in movies.

```
#created box plot for Genre against Scores

boxplot(Score ~ Genre, data = data_movie, col = c("red","green", "yellow"))
```



From the plot, it can be seen that while Action and Comedy appear to be somewhat closely varied, Drama has a bigger variation than both Action and Comedy in terms of the number of brands recalled. Moreover, despite drama appearing to be a little bit skewed, the three Genres generally are close to being normally distributed. These differences, could indicate that genre is also an important factor when it comes to brand recall in movies.

```r
#Created the comparative box plot for Gender and Genre against Scores

boxplot(Score ~ Gender + Genre, data = data_movie, col = c("pink","blue"))
```



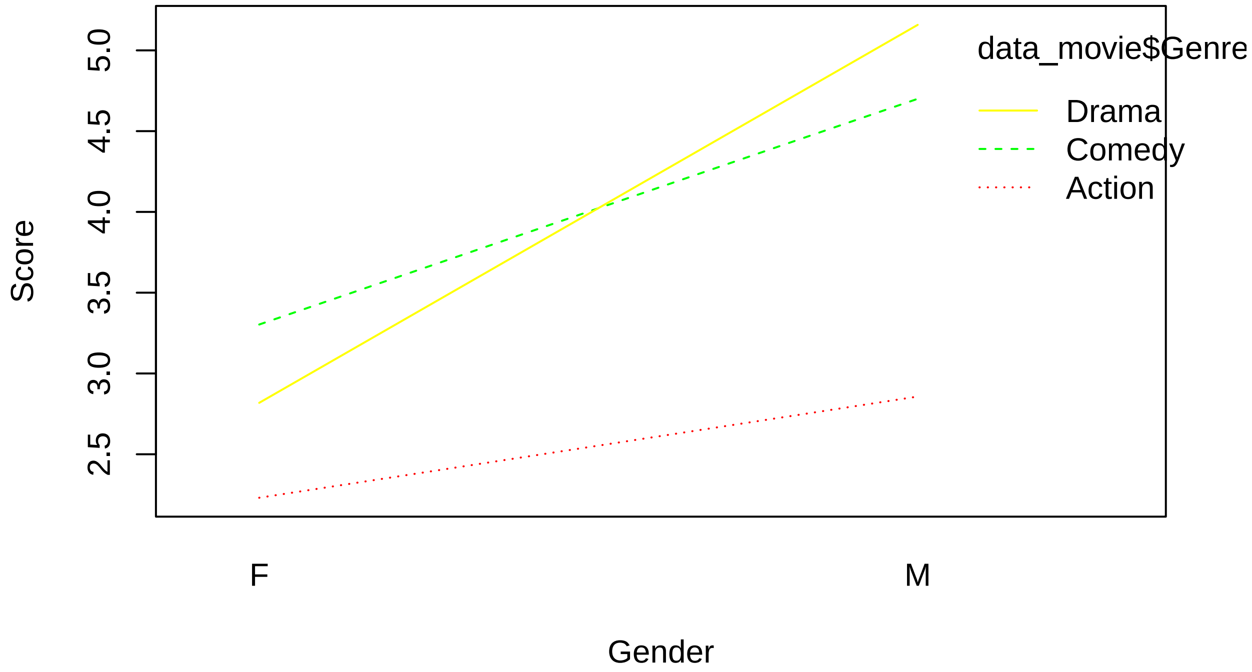
From the plot, it can be seen that the combination of the two factors, namely, Gender and Genre have varying effects. It appears that males who watched drama and comedy, have recalled the most brands overall. Contrary to males, females who watched drama appeared to have low recall scores. Similarly, females who watched comedy seemed to recall the most brands for females.

To have a clearer understanding of the different interactions among the two factors of the data, and the response; I will be constructing interaction plots as well.

Interaction Plot

Considering that we have two factors, it is best to use the interaction plot to see if the response changes along with the changes in the different levels of the two factors.

```
#plotted the interaction plot
interaction.plot(data_movie$Gender, data_movie$Genre, data_movie$Score, xlab = "Gender", ylab = "Score"
```



From the plot, it can be seen that the change in response to changing the Gender factor becomes different as the Genre factor levels change. It is clear from the plot that the slopes of the lines are different and are not parallel from each other which means that there are interactions among these factors.Based on the figure, combined male gender and drama genre together significantly increases the effect on the scores of brand recall as compared to other combinations.

Given that this is an unbalanced study with two factors, and given that we are interested to include the interaction term (if there are any significant) between these two factors; the full mathematical model for this situation is:

$$Yi = \mu + \alpha i + \beta j + \gamma ij + \epsilon ijk$$

where:

Yi is the Score response;

μ is the overall mean or intercept;

αi is the main gender effect for i = 1, 2;

βj is the main genre effect for j = 1, 2, 3;

γ ij is the interaction effect between gender and genre;

ϵ ijk is the random error term or residual term, which represents the unexplained variation in the response variable

Since the study is unbalanced, the order will matter when constructing ANOVA tables. From the box plot and the interaction graph above, it is evident that both Genre and Gender have an effect on growth. To ensure that each factor has an effect, I will first conduct a One-way ANOVA test on both factors.

One-way ANOVA on Gender

The following are the null and alternative hypothesis tests for the One-way ANOVA on Gender.

H0: $\alpha 1 = \alpha 2 = 0$

This null hypothesis assumes that there is no significant difference in the mean scores or coefficients across the different levels of Gender or in other words, Gender has no significant effect on the score.

H1: At least one $\alpha i \neq 0$

The alternative hypothesis assumes that there is at least one level of the Gender factor that has a significant difference in the mean scores across the different levels of Gender or in other words, Gender has a significant effect on the Score.

```r
#conducted One-Way ANOVA
one_way_gender = lm(Score ~ Gender, data = data_movie)
anova(one_way_gender)
```

```
## Analysis of Variance Table
##
## Response: Score
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## Gender      1  71.583  71.583  52.824 2.648e-11 ***
## Residuals 135 182.942   1.355
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Based on the result, the p-value is 2.648e-11, which is a lot smaller than 0.05. Therefore, we can reject the null hypothesis which means that *Gender* has a *significant effect on Score*, thus, validating my preliminary analysis earlier.

One-way ANOVA on Genre

The following are the null and alternative hypothesis tests for the One-way ANOVA on Genre.

H0: $\beta 1 = \beta 2 = \beta 3 = 0$
This null hypothesis assumes that there is no significant difference in the mean scores or coefficients across the different levels of Genre or in other words, Genre has no significant effect on the score.

H1: At least one $\beta i \neq 0$
The alternative hypothesis assumes that there is at least one level of the Genre factor that has a significant difference in the mean scores across the different levels of Genre or in other words, Genre has a significant effect on the Score.

```r
#conducted One-Way ANOVA
one_way_genre = lm(Score ~ Genre, data = data_movie)
anova(one_way_genre)
```

```
## Analysis of Variance Table
##
## Response: Score
##            Df Sum Sq Mean Sq F value    Pr(>F)
## Genre       2  62.19 31.0950  21.664 7.047e-09 ***
## Residuals 134 192.34  1.4353
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Similarly, based on the result, the p-value is 7.047e-09, which is also a lot smaller than 0.05. Therefore, we can also reject the null hypothesis which means that Genre is also significant on Score, thus, also validating my preliminary analysis earlier.

Two-way ANOVA

The following are the null and alternative hypothesis tests for the Two-way ANOVA.

H0: $\alpha i1 + \alpha i2 + \beta j1 + \beta j2 + \beta j3 = 0$; for all i, j

The null hypothesis assumes that the sum of the effects of all levels of Gender and Genre is equal to zero, which means that there are no significant main effects. This means that there is no interaction effect between Gender and Genre on the Score, hence, the reason why there is no interaction term included in the test as this solely focuses on the main effects of Gender and Genre as separate factors.

H1: At least one αi or $\beta j \neq 0$
The alternative hypothesis suggests that at least one of the levels of Gender or Genre is not equal to zero, which means that at least one level of Gender or Genre has a significant effect on Score.

```
#conducted Two-way ANOVA
two_way_gender = lm(Score ~ Gender + Genre, data = data_movie)
anova(two_way_gender)
```

```
## Analysis of Variance Table
##
## Response: Score
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## Gender      1  71.583  71.583  71.807 3.914e-14 ***
## Genre       2  50.357  25.178  25.257 5.036e-10 ***
## Residuals 133 132.585   0.997
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the Two-way ANOVA conducted, the p-values of both Gender and Genre are below 0.05 which means that we can reject the null hypothesis, hinting that at least one of the gender or genre levels are significant.

Two-way ANOVA with Interaction Term

For the next part, I will be conducting a Two-way ANOVA with an interaction term and determine if the interaction term is significant for the model.

The following are the null and alternative hypothesis tests for the Two-way ANOVA with interaction term:

H0 : $\gamma ij = 0$; for all i, j

The null hypothesis assumes that there is no interaction effect between the Gender and Genre variables on the Score.

H1 : at least one $\gamma ij \neq 0$

The alternative hypothesis assumes that there is an interaction effect between the Gender and Genre variables on the Score.

```
#got summaries for both separate One-way tests conducted to get the coefficients and p-values
summary(one_way_gender)
```

```
##
## Call:
## lm(formula = Score ~ Gender, data = data_movie)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3023 -0.7447  0.2553  0.6977  2.2553
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.7447     0.1201  22.859  < 2e-16 ***
## GenderM       1.5576     0.2143   7.268 2.65e-11 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.164 on 135 degrees of freedom
## Multiple R-squared:  0.2812, Adjusted R-squared:  0.2759
## F-statistic: 52.82 on 1 and 135 DF,  p-value: 2.648e-11
```

```
summary(one_way_genre)
```

```
##
## Call:
## lm(formula = Score ~ Genre, data = data_movie)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.90244 -0.62791  0.09756  0.60377  2.60377
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.3962     0.1646  14.561  < 2e-16 ***
## GenreComedy   1.2317     0.2459   5.009 1.69e-06 ***
## GenreDrama    1.5062     0.2492   6.045 1.40e-08 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.198 on 134 degrees of freedom
## Multiple R-squared:  0.2443, Adjusted R-squared:  0.2331
## F-statistic: 21.66 on 2 and 134 DF,  p-value: 7.047e-09
```

```
#conducted Two-way ANOVA with Interaction Term
two_way_int = lm(Score ~ Gender * Genre, data = data_movie)
anova(two_way_int)
```

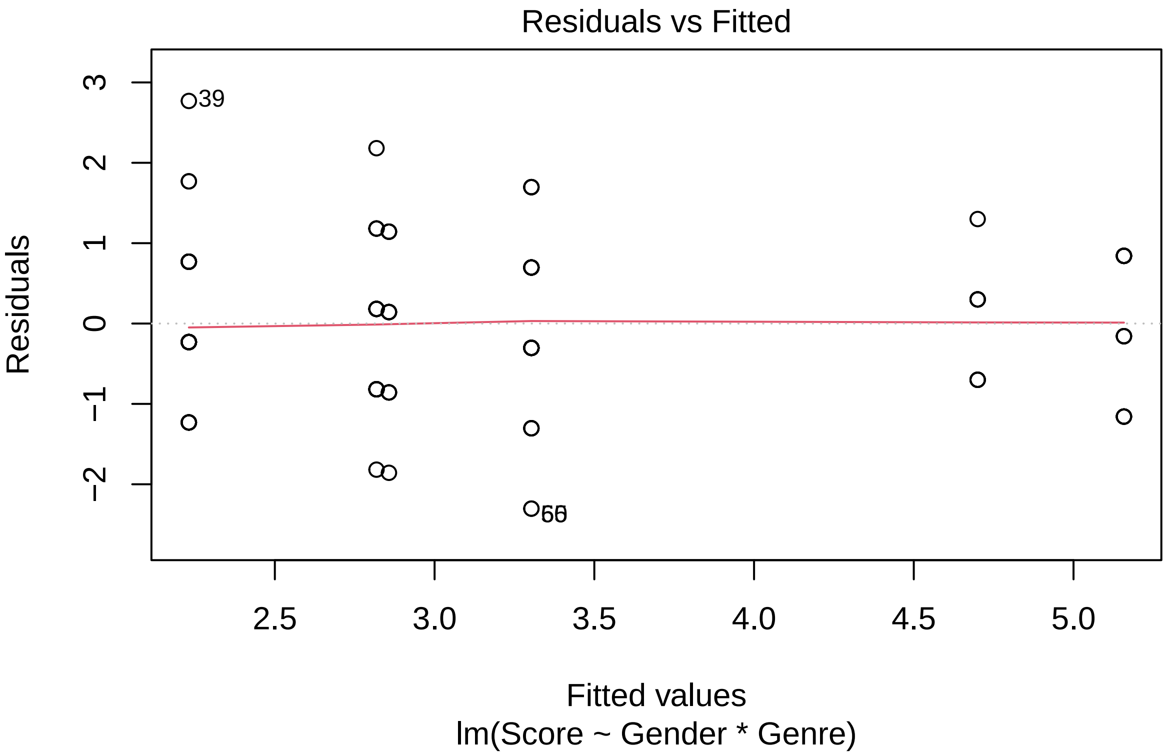
```
## Analysis of Variance Table
##
## Response: Score
##               Df  Sum Sq Mean Sq F value    Pr(>F)
## Gender         1  71.583  71.583 79.8038 3.277e-15 ***
## Genre          2  50.357  25.178 28.0698 7.152e-11 ***
## Gender:Genre   2  15.079   7.540  8.4054 0.0003677 ***
## Residuals    131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

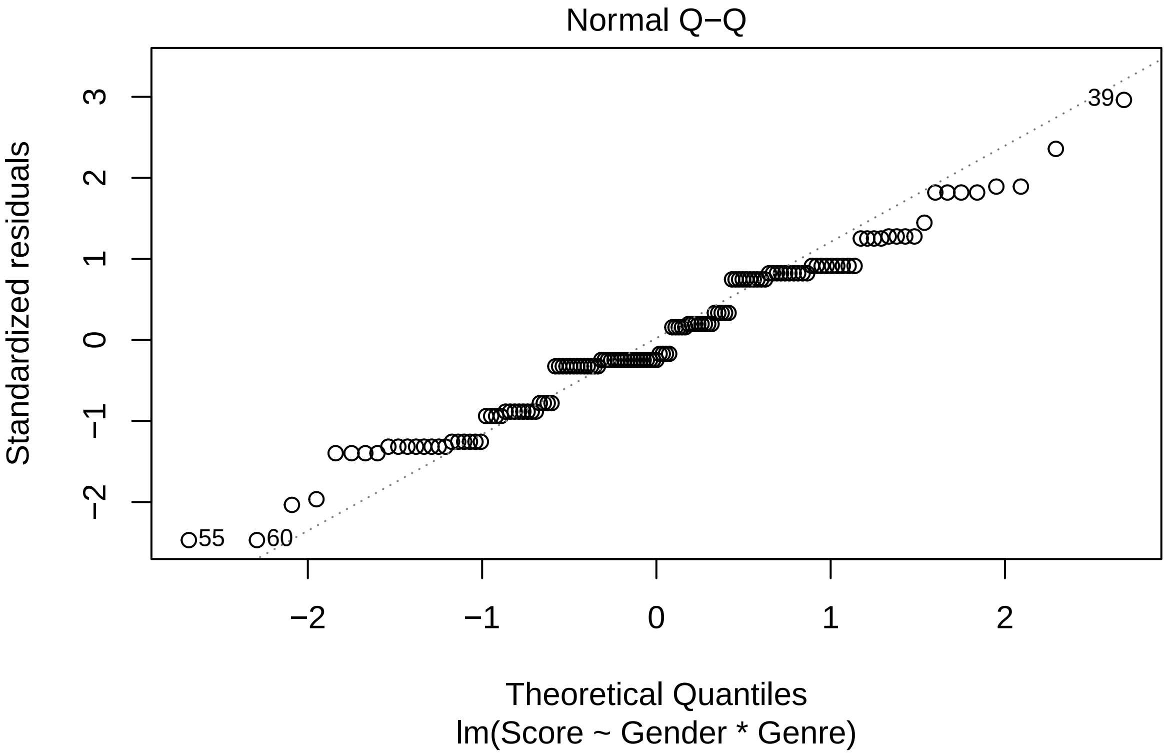
Based on the ANOVA table, the p-value of the interaction term, 0.0003677, is less than 0.05 which means that we can reject the null hypothesis claiming that there is no interaction effect between the Gender and Genre factors. In other words, the interaction term is considered to be significant. Therefore, the interaction term will be kept and the mathematical model to be used as basis will remain as-is where both factors Gender and Gender will be used along with the Interaction Term.

Considering that factors with larger coefficients and smaller p-values suggest stronger associations and greater significant effects, the model to be used moving forward is still the model where Gender is fitted first because based on the one-way ANOVA conducted previously for both Gender and Genre; Gender has the lowest p-value; and based on the two-way ANOVA conducted, Gender still has the lowest p-value, which has a huge difference against the Genre p-value.

Model Validation

```
#mapped the diagnostic plots for validation
plot(two_way_int, which = 1:2)
```


Residuals vs Fitted
Residuals
3
2
1
0
−1
−2
39
57
60
2.5
3.0
3.5
4.0
4.5
5.0
Fitted values
lm(Score ~ Gender * Genre)



Based on the Residuals vs Fitted plot, given that the residuals are evenly scattered around the residual line where y = 0. In this figure, there is no specific trend or pattern, which is an indication that the model's assumptions are met. Moreover, the normal quantile plot of residuals appear to be closer to linear suggesting residuals are close to normally distributed, which is another assumption that has been met.

Now that the assumptions of the model have been validated, we can safely say that the *interaction term* (Gender * Genre) in the model is statistically significant, indicating the presence of an interaction effect.

Based on the analysis activities performed, it can be concluded that the drama genre has a significant effect on the brand recall score. Just by looking at the preliminary investigation graphs, it is evident that individuals who watched drama movies had higher brand recall scores compared to those who watched other genres. Moreover, the final selected model supports the assumption that the interaction term is deemed to be significant, which means that the combination of Gender and Genre has a significant effect on brand recognition. With this said, it is important to recall that based on the preliminary graphs, males who watched drama movies had the highest scores. Therefore, to maximize brand placements, I am recommending for the business to have brand placements on *drama* movies catered to the *male* audience.