Probability and Statistics for Data Analysis

ASSIGNMENT 2

Stylianos Giagkos | f3352410

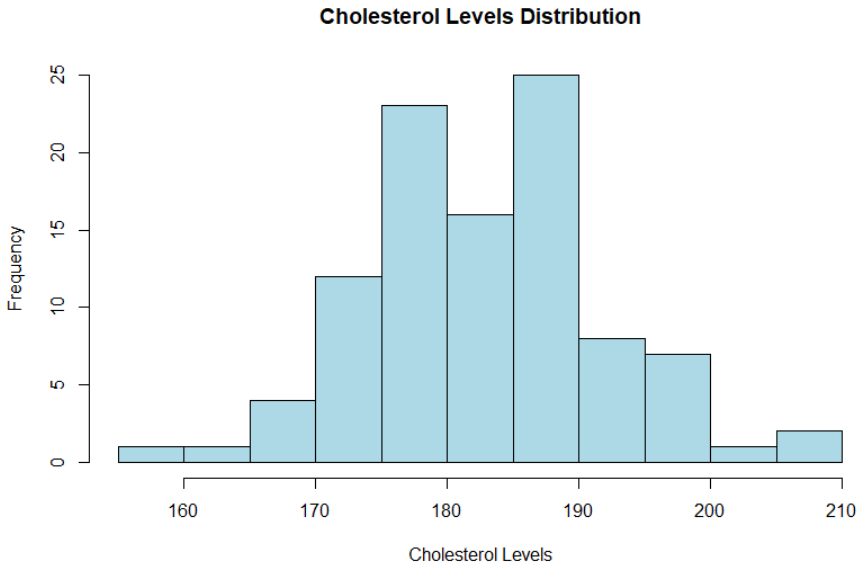
Probability and Statistics for Data Analysis
ASSIGNMENT 2
Stylianos Giagkos | f3352410

1. A study of the effect of two drugs on the reduction of cholesterol used 100 volunteers who tested the drugs. Fifty of them were randomly selected to take the first drug (A), while the remaining fifty took the second one (B). We measured the cholesterol levels after receiving the drugs, tested the presence of Myalgia symptoms, and measured the Glucose levels in order to check the side effects of the drugs. The observations are in the 1 file "cholesterol.txt"

(a) Provide a 99% confidence interval for Cholesterol values

```
# Load data
cholesterol_data <- read.table("cholesterol.txt", header = TRUE)

# 1. (a) 99% Confidence Interval for Cholesterol values
cholesterol_values <- cholesterol_data$Cholesterol
mean_cholesterol <- mean(cholesterol_values)
std_error <- sd(cholesterol_values) / sqrt(length(cholesterol_values))

# Calculate 99% Confidence Interval
z_score <- qnorm(0.995) # for 99% confidence
ci_lower <- mean_cholesterol - z_score * std_error
ci_upper <- mean_cholesterol + z_score * std_error
cat("99% Confidence Interval for Cholesterol: [", ci_lower, ",", ci_upper, "]\n")

# Plot for Cholesterol values
hist(cholesterol_values, main = "Cholesterol Levels Distribution", xlab = "Cholesterol Levels", col =
"lightblue", border = "black")
```

99% Confidence Interval for Cholesterol: [180.73, 185.47]
This confidence interval means we are 99% confident that the true mean cholesterol value for the population lies between 180.73 and 185.47.

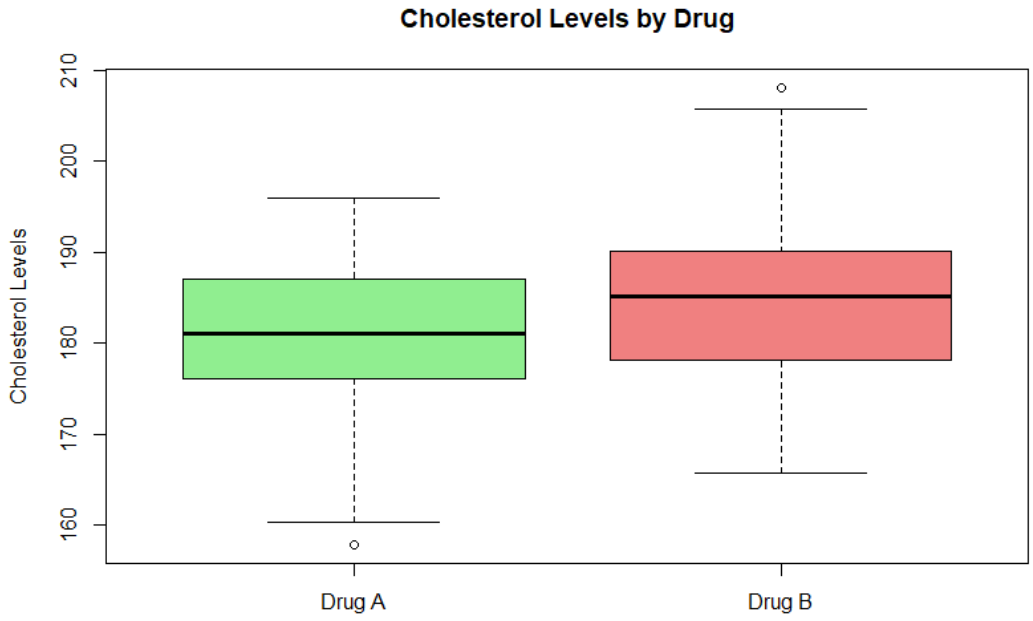


(b) Provide a 95% confidence interval for Cholesterol values after receiving drug A and B, respectively

```
# 2. (b) 95% Confidence Interval for Cholesterol values for drugs A and B
cholesterol_A <- cholesterol_data$Cholesterol[cholesterol_data$Drug == "A"]
cholesterol_B <- cholesterol_data$Cholesterol[cholesterol_data$Drug == "B"]

# Calculate 95% CI for Cholesterol values after receiving drug A and B
ci_A <- t.test(cholesterol_A, conf.level = 0.95)$conf.int
ci_B <- t.test(cholesterol_B, conf.level = 0.95)$conf.int
cat("95% Confidence Interval for Cholesterol (Drug A):", ci_A, "\n")
cat("95% Confidence Interval for Cholesterol (Drug B):", ci_B, "\n")

# Plot for Cholesterol levels by Drug
boxplot(cholesterol_A, cholesterol_B, names = c("Drug A", "Drug B"), main = "Cholesterol Levels by Drug", ylab
= "Cholesterol Levels", col = c("lightgreen", "lightcoral"))
```

95% Confidence Interval for Cholesterol (Drug A): [178.66, 183.37]
This confidence interval suggests that the mean cholesterol level for those taking Drug A lies between 178.66 and 183.37 with 95% confidence.

95% Confidence Interval for Cholesterol (Drug B): [182.43, 187.94]
The mean cholesterol level for those taking Drug B is expected to lie between 182.43 and 187.94 with 95% confidence.

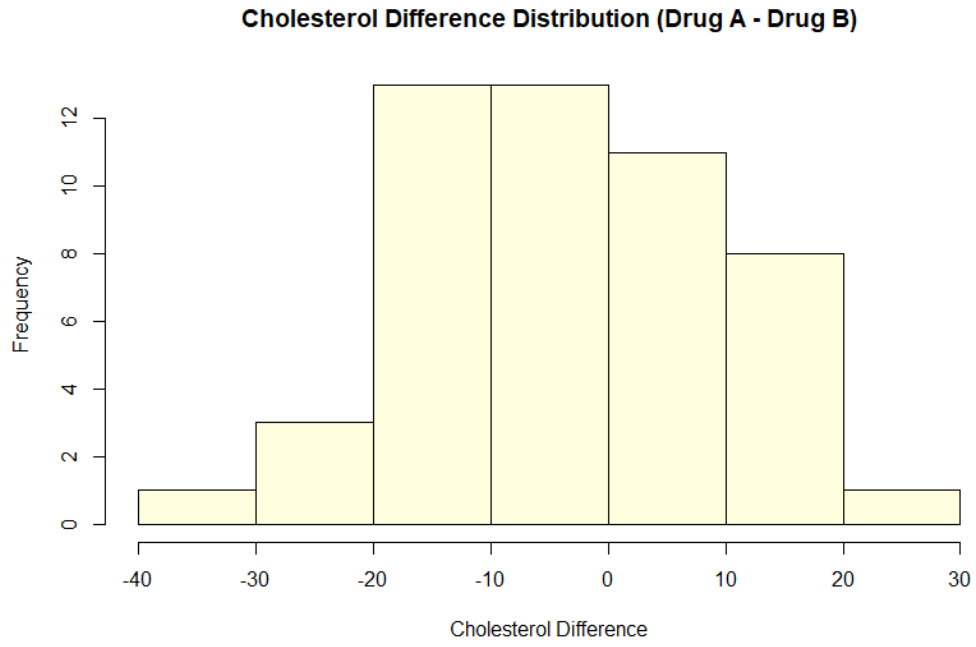


(c) Provide a 90% confidence interval for the mean difference in Cholesterol values after receiving drug A and drug B, respectively.

```
# 3. (c) 90% Confidence Interval for the mean difference in Cholesterol values
cholesterol_diff <- cholesterol_A - cholesterol_B
mean_diff <- mean(cholesterol_diff)
std_error_diff <- sd(cholesterol_diff) / sqrt(length(cholesterol_diff))

# Calculate Confidence Interval
z_score_90 <- qnorm(0.95) # for 90% confidence
ci_diff_lower <- mean_diff - z_score_90 * std_error_diff
ci_diff_upper <- mean_diff + z_score_90 * std_error_diff
cat("90% Confidence Interval for the mean difference in Cholesterol: [", ci_diff_lower, ",", ci_diff_upper,
"]\n")

# Plot for Cholesterol difference
hist(cholesterol_diff, main = "Cholesterol Difference Distribution (Drug A - Drug B)", xlab = "Cholesterol
Difference", col = "lightyellow", border = "black")
```

90% Confidence Interval for the Mean Difference in Cholesterol (Drug A - Drug B): [-7.19, -1.15]

The confidence interval for the mean difference between cholesterol levels in Drug A and Drug B is negative, suggesting that Drug A generally results in lower cholesterol levels compared to Drug B. With 90% confidence, the true mean difference lies between -7.19 and -1.15.

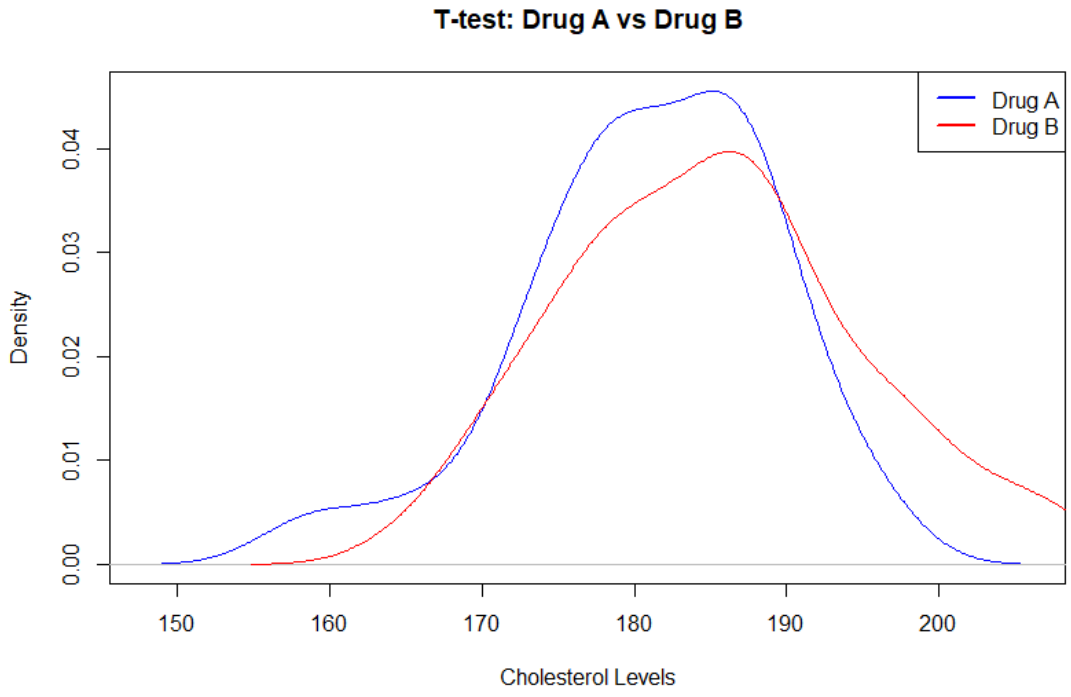


(d) Examine the following hypothesis test: Ho : μA = μB H1 : μA < μB where μA and μB are the mean Cholesterol values after receiving drug A and B, respectively. The level of significance is α = 0.05.

```
# 4. (d) Hypothesis Test for H0: μA = μB, H1: μA < μB
t_test_result <- t.test(cholesterol_A, cholesterol_B, alternative = "less")
cat("Hypothesis Test (p-value):", t_test_result$p.value, "\n")

# Plot for Hypothesis Test (T-test result)
plot(density(cholesterol_A), main = "T-test: Drug A vs Drug B", xlab = "Cholesterol
Levels", col = "blue")
lines(density(cholesterol_B), col = "red")
legend("topright", legend = c("Drug A", "Drug B"), col = c("blue", "red"), lwd = 2)
```

T-test for Hypothesis (p-value): 0.0114

The p-value of 0.0114 indicates that there is significant evidence to reject the null hypothesis (that the mean cholesterol levels for Drug A and Drug B are equal) in favor of the alternative hypothesis that Drug A leads to lower cholesterol.



(e) Provide a hypothesis test (α = 0.01) for the equality of variances of Glucose levels after receiving drug A and drug B, respectively

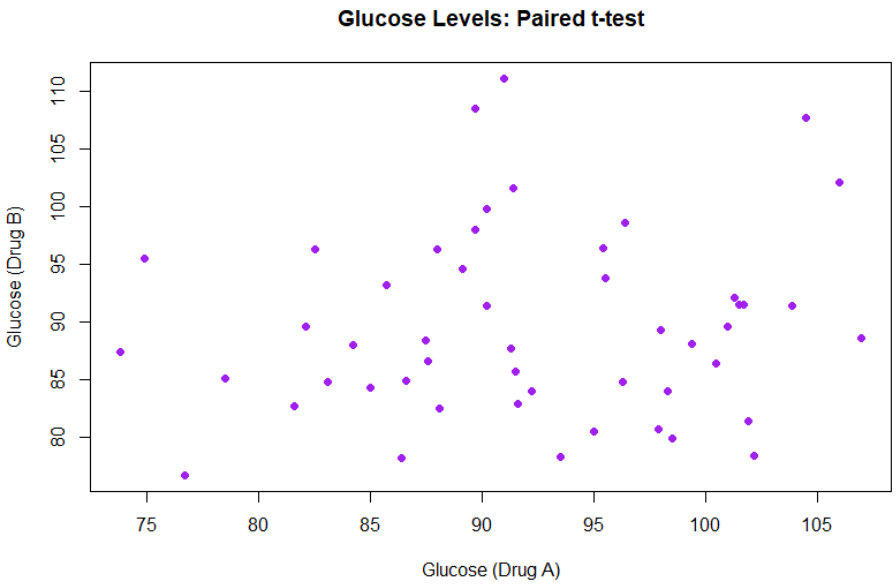
```
# 5. (e) Test for Equality of Variances of Glucose Levels
glucose_A <- cholesterol_data$Glucose[cholesterol_data$Drug == "A"]
glucose_B <- cholesterol_data$Glucose[cholesterol_data$Drug == "B"]

var_test_result <- var.test(glucose_A, glucose_B, alternative = "two.sided")
cat("Test for Equality of Variances (p-value):", var_test_result$p.value, "\n")
```

Equality of Variances Test (p-value): 0.8694
The p-value is high (0.8694), indicating that there is no significant difference in the variances of glucose levels between those on Drug A and Drug B. This suggests that the variability of glucose levels between the two groups is similar.

(f) At a significance level of 5%, test if there is a statistically significant side effect on Glucose levels

```
# 6. (f) Test for statistically significant side effect on Glucose levels
paired_t_test_result <- t.test(glucose_A, glucose_B, paired = TRUE)
cat("Paired t-test for Glucose levels (p-value):", paired_t_test_result$p.value,
"\n")
# Plot for Paired t-test (Glucose levels)
plot(glucose_A, glucose_B, main = "Glucose Levels: Paired t-test", xlab = "Glucose
(Drug A)", ylab = "Glucose (Drug B)", col = "purple", pch = 16)
```

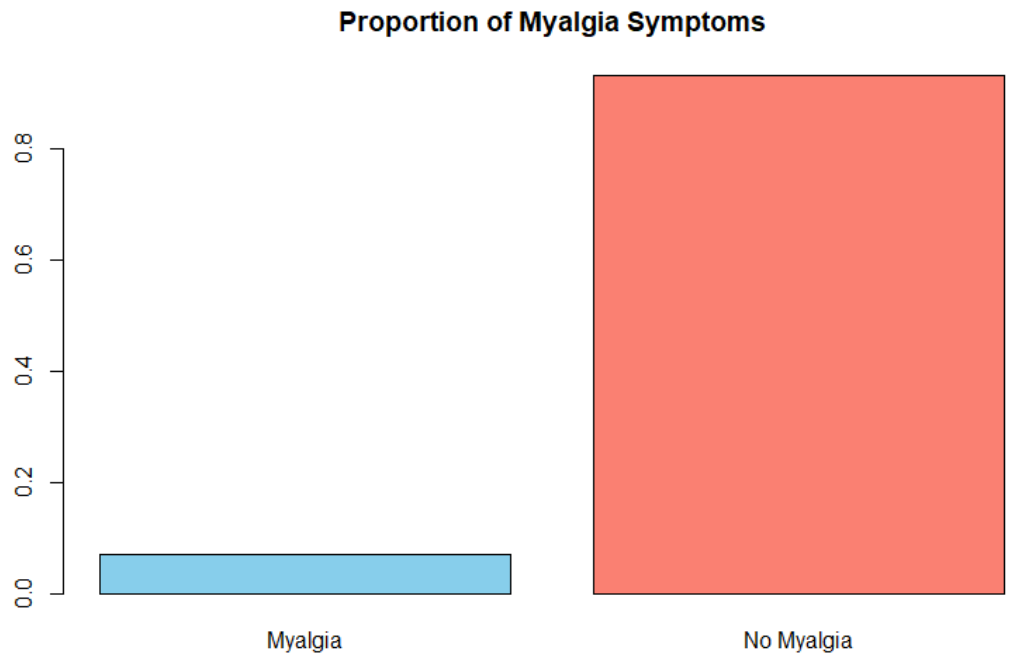
Paired T-test for Glucose levels (p-value): 0.1117
This p-value suggests that there is no significant difference in glucose levels between those who used Drug A and those who used Drug B when tested as paired data.



(g) Provide a 95% confidence interval for the proportion of volunteers who had Myalgia symptoms.

```
# 7. (g) 95% Confidence Interval for Proportion of Myalgia Symptoms
myalgia_data <- cholesterol_data$Myalgia
myalgia_yes <- sum(myalgia_data == "Yes")
myalgia_no <- sum(myalgia_data == "No")
proportion_myalgia <- myalgia_yes / (myalgia_yes + myalgia_no)

# Confidence Interval for Proportion
se_proportion <- sqrt(proportion_myalgia * (1 - proportion_myalgia) / (myalgia_yes +
myalgia_no))
ci_proportion_lower <- proportion_myalgia - qnorm(0.975) * se_proportion
ci_proportion_upper <- proportion_myalgia + qnorm(0.975) * se_proportion
cat("95% Confidence Interval for Proportion of Myalgia Symptoms: [",
ci_proportion_lower, ",", ci_proportion_upper, "]\n")

# Plot for Myalgia Proportion
barplot(c(proportion_myalgia, 1 - proportion_myalgia), names.arg = c("Myalgia", "No
Myalgia"), main = "Proportion of Myalgia Symptoms", col = c("skyblue", "salmon"))
```

95% Confidence Interval for Proportion of Myalgia Symptoms: [0.02, 0.12]
This confidence interval shows that with 95% confidence, the proportion of individuals experiencing myalgia symptoms lies between 2% and 12%.



(h) Test if the proportion of volunteers who had Myalgia symptoms is statistically greater than 5% at a significance level of 5%.

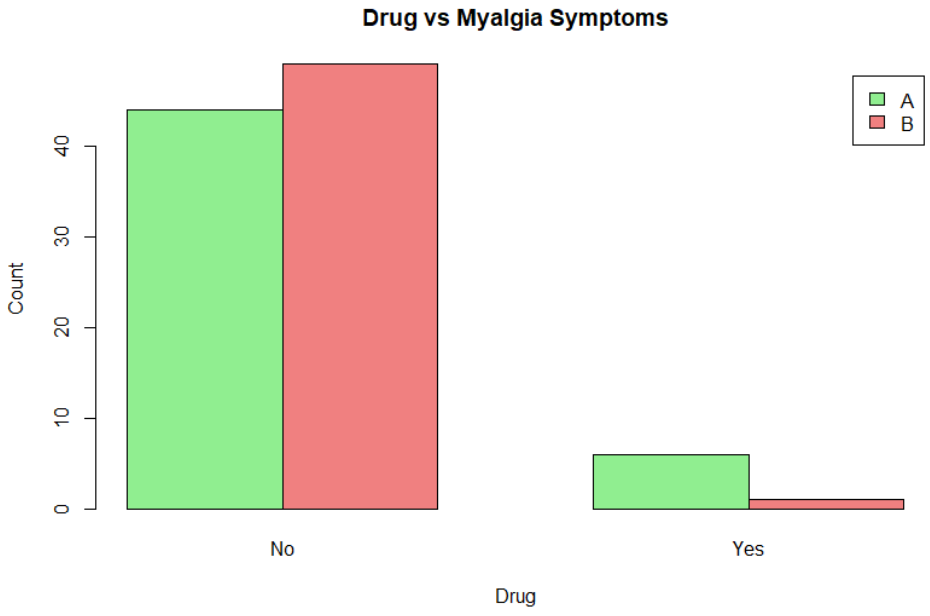
```
# 8. (h) Hypothesis Test for Myalgia Proportion Greater than 5%
prop_test_result <- prop.test(myalgia_yes, myalgia_yes + myalgia_no, alternative = "greater", p = 0.05)
cat("Hypothesis Test for Myalgia Proportion > 5% (p-value):", prop_test_result$p.value, "\n")
```

Hypothesis Test for Myalgia Proportion > 5% (p-value): 0.2456
The p-value of 0.2456 suggests that there is no significant evidence to conclude that the proportion of myalgia symptoms is greater than 5%.

(i) Test if the drug and the presence of Myalgia symptoms are independent (α = 0.05).

```
# 9. (i) Test for Independence between Drug and Myalgia Symptoms
table_drug_myalgia <- table(cholesterol_data$Drug, cholesterol_data$Myalgia)

# Fisher's Exact Test for Independence between Drug and Myalgia Symptoms
fisher_test_result <- fisher.test(table_drug_myalgia)
cat("Fisher's Exact Test for Independence (p-value):", fisher_test_result$p.value, "\n")

# Plot for Fisher's Exact Test (Drug vs Myalgia Symptoms)
barplot(table_drug_myalgia, beside = TRUE, col = c("lightgreen", "lightcoral"),
legend = TRUE, main = "Drug vs Myalgia Symptoms", xlab = "Drug", ylab = "Count")
```

Fisher's Exact Test for Independence (p-value): 0.1117
This p-value suggests that there is no significant association between drug type and the occurrence of myalgia symptoms. This means that the likelihood of experiencing myalgia does not significantly depend on whether the person used Drug A or Drug B.

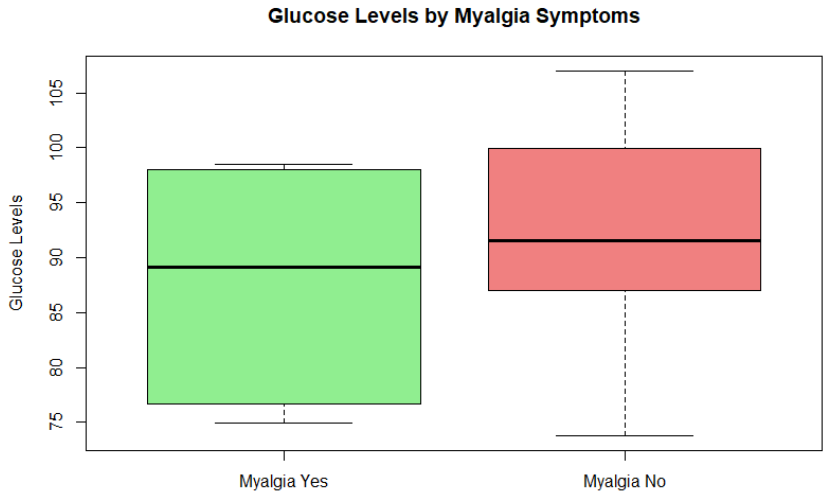


(j) Provide a 95% confidence interval for the mean difference μ1 – μ2, where μ1 and μ2 are the mean Glucose levels for volunteers with and without Myalgia symptoms, respectively.

```
# 10. (j) 95% Confidence Interval for Mean Difference in Glucose Levels (Myalgia Yes vs No)
glucose_myalgia_yes <- glucose_A[myalgia_data == "Yes"]
glucose_myalgia_no <- glucose_A[myalgia_data == "No"]

#Use t-test to compare the two groups' glucose levels:
t_test_glucose_result <- t.test(glucose_myalgia_yes, glucose_myalgia_no)
cat("T-test for Mean Difference in Glucose Levels (p-value):",
t_test_glucose_result$p.value, "\n")

# 95% Confidence Interval for the mean difference
ci_glucose_lower <- t_test_glucose_result$conf.int[1]
ci_glucose_upper <- t_test_glucose_result$conf.int[2]
cat("95% Confidence Interval for Mean Difference in Glucose Levels (Myalgia Yes vs No): [", ci_glucose_lower, ",", ci_glucose_upper, "]\n")

# Plot for Glucose Levels Difference
boxplot(glucose_myalgia_yes, glucose_myalgia_no, names = c("Myalgia Yes", "Myalgia No"), main = "Glucose Levels by Myalgia Symptoms", ylab = "Glucose Levels", col = c("lightgreen", "lightcoral"))
```

T-test for Mean Difference in Glucose Levels (p-value): 0.2924
This p-value indicates that there is no significant difference in glucose levels between people with and without myalgia symptoms.

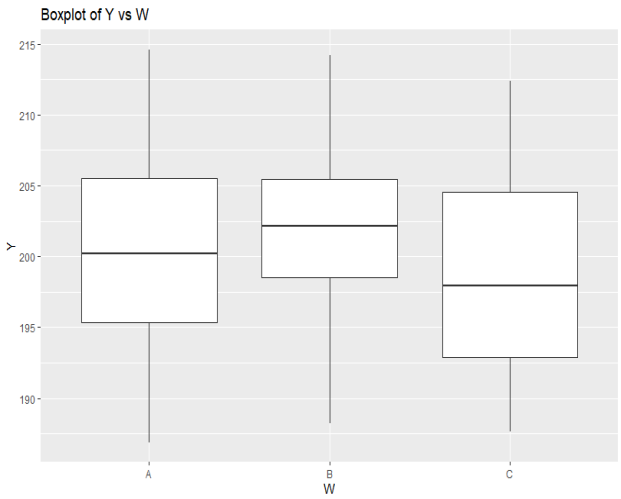
95% Confidence Interval for Mean Difference in Glucose Levels (Myalgia Yes vs No): [-15.58, 5.62]
The confidence interval for the mean difference in glucose levels between those with and without myalgia symptoms is wide and includes zero. This suggests that there is no clear difference in glucose levels based on whether or not someone experiences myalgia symptoms.

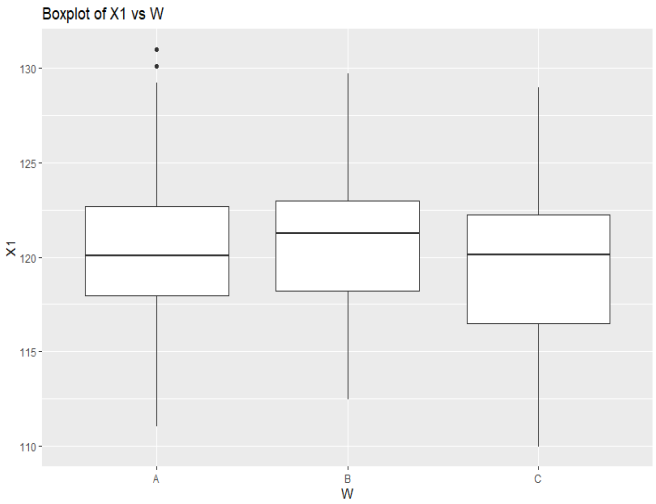


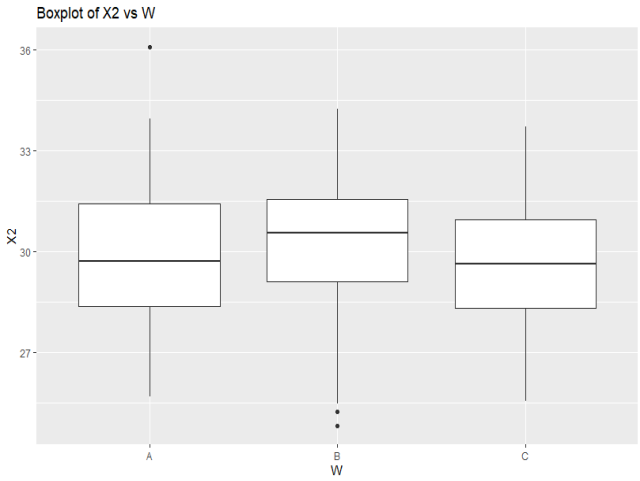
2. In the file "data2.txt" (available on the e-class assignments site), you will find the recorded variables Y, X1, X2, X3, X4 (continuous), and W (categorical with three levels) for 150 cases. Using these data, answer the following questions:

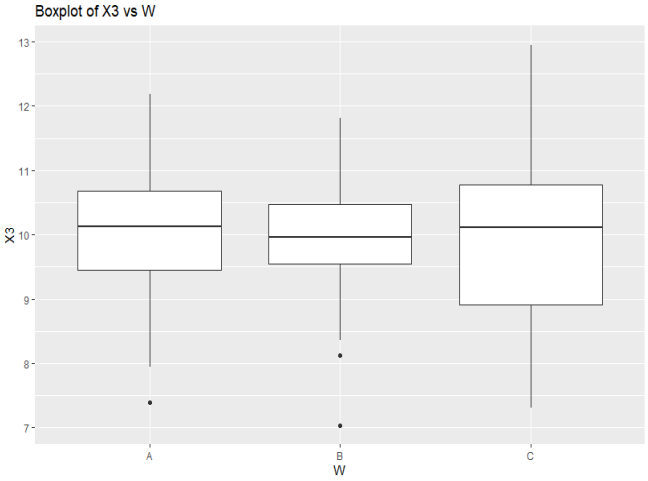
(a) Run the parametric one-way ANOVA for each of the continuous variables (Y, X1, X2, X3, X4) on the categorical variable (W). Specifically,

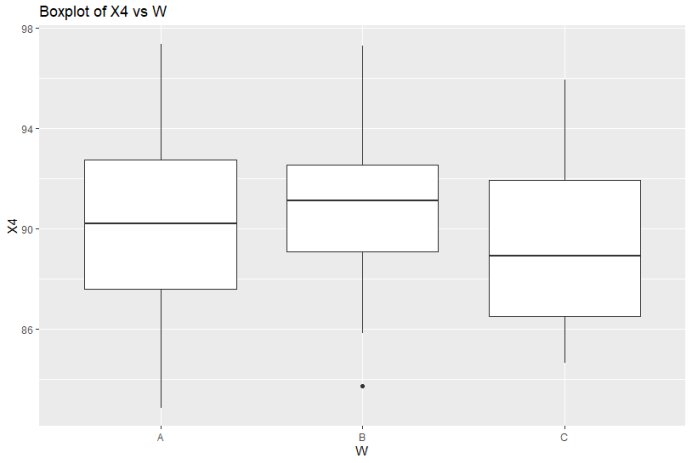
(i) Provide a graphical representation of each continuous variable versus the categorical variable.











(ii) Provide the ANOVA output.

```
ANOVA results for Y
> summary(anova_Y)
             Df Sum Sq Mean Sq F value Pr(>F)
W             2    333  166.71   4.352 0.0141 *
Residuals   197   7546   38.31
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The ANOVA result indicates that the categorical variable W does significantly affect the continuous variable Y, as the p-value is 0.0141, which is less than the 0.05 significance level.
- The F value of 4.352 indicates that there is a moderate to strong relationship between the levels of W and the variation in Y.
- Since the p-value is less than 0.05, we conclude that at least one of the group means of Y is different from the others, suggesting that W has a significant impact on Y.

```
ANOVA results for X1
> summary(anova_X1)
             Df Sum Sq Mean Sq F value Pr(>F)
W             2   76.3   38.13    2.42 0.0915 .
Residuals   197 3104.1   15.76
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The ANOVA result suggests that while there is some variation in X1 across the levels of W, the evidence is not statistically significant at the 5% significance level ($p = 0.0915 > 0.05$).
- The F value of 2.42 indicates a moderate amount of variation explained by W, but since the p-value is above 0.05, it is not statistically significant. Therefore, we conclude that the categorical variable W does not have a significant impact on the continuous variable X1 in this case.

```
ANOVA results for X2
> summary(anova_X2)
             Df Sum Sq Mean Sq F value Pr(>F)
W             2   17.0   8.489   2.079  0.128
Residuals   197  804.3   4.083
```

- The ANOVA result suggests that the categorical variable W **does not significantly affect** the continuous variable X2, as the p-value is 0.128, which is greater than the commonly used significance level of 0.05.
- The F value of 2.079 indicates a moderate amount of variation explained by W, but the p-value being above 0.05 means that we do not have enough evidence to conclude that W has a significant effect on X2.
- Therefore, **W does not have a statistically significant impact on X2** in this case.

```
ANOVA results for X3
> summary(anova_X3)
             Df Sum Sq Mean Sq F value Pr(>F)
W             2   0.28  0.1397   0.133  0.876
Residuals   197 207.24  1.0520
```

- The ANOVA result suggests that the categorical variable W does not significantly affect the continuous variable X3, as the p-value is 0.876, which is much greater than the 0.05 significance level.
- The F value of 0.133 indicates very little variation explained by W, and the high p-value further confirms that W does not have a significant impact on X3 in this case.

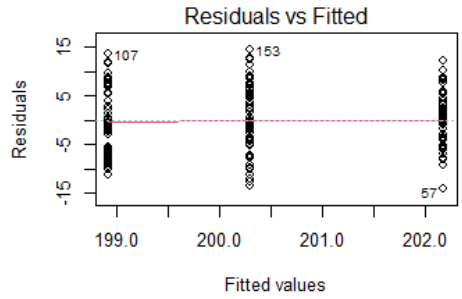
```
ANOVA results for X4
> summary(anova_X4)
             Df Sum Sq Mean Sq F value Pr(>F)
W             2   75.8   37.89   4.171 0.0168 *
Residuals   197 1789.6    9.08
```

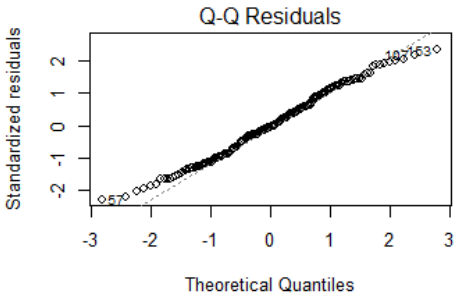
- The ANOVA result indicates that the categorical variable W does significantly affect the continuous variable X4, as the p-value is 0.0168, which is less than the 0.05 significance level.
- The F value of 4.171 indicates a moderate to strong relationship between the levels of W and the variation in X4.
- Since the p-value is below 0.05, we conclude that there is a significant difference in the means of X4 across the levels of W.

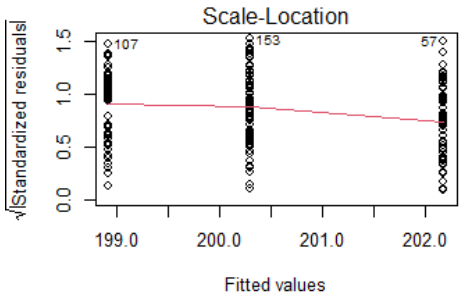
– **Conclusions for All Variables:**

- **For Y: There is a statistically significant effect of W on Y (p-value = 0.0141 < 0.05).**
- **For X1**: There is no statistically significant effect of W on X1 (p-value = 0.0915 > 0.05).
- **For X2**: There is no statistically significant effect of W on X2 (p-value = 0.128 > 0.05).
- **For X3**: There is no statistically significant effect of W on X3 (p-value = 0.876 > 0.05).
- **For X4**: There is a statistically significant effect of W on X4 (p-value = 0.0168 < 0.05).

(iii) Check the assumptions.



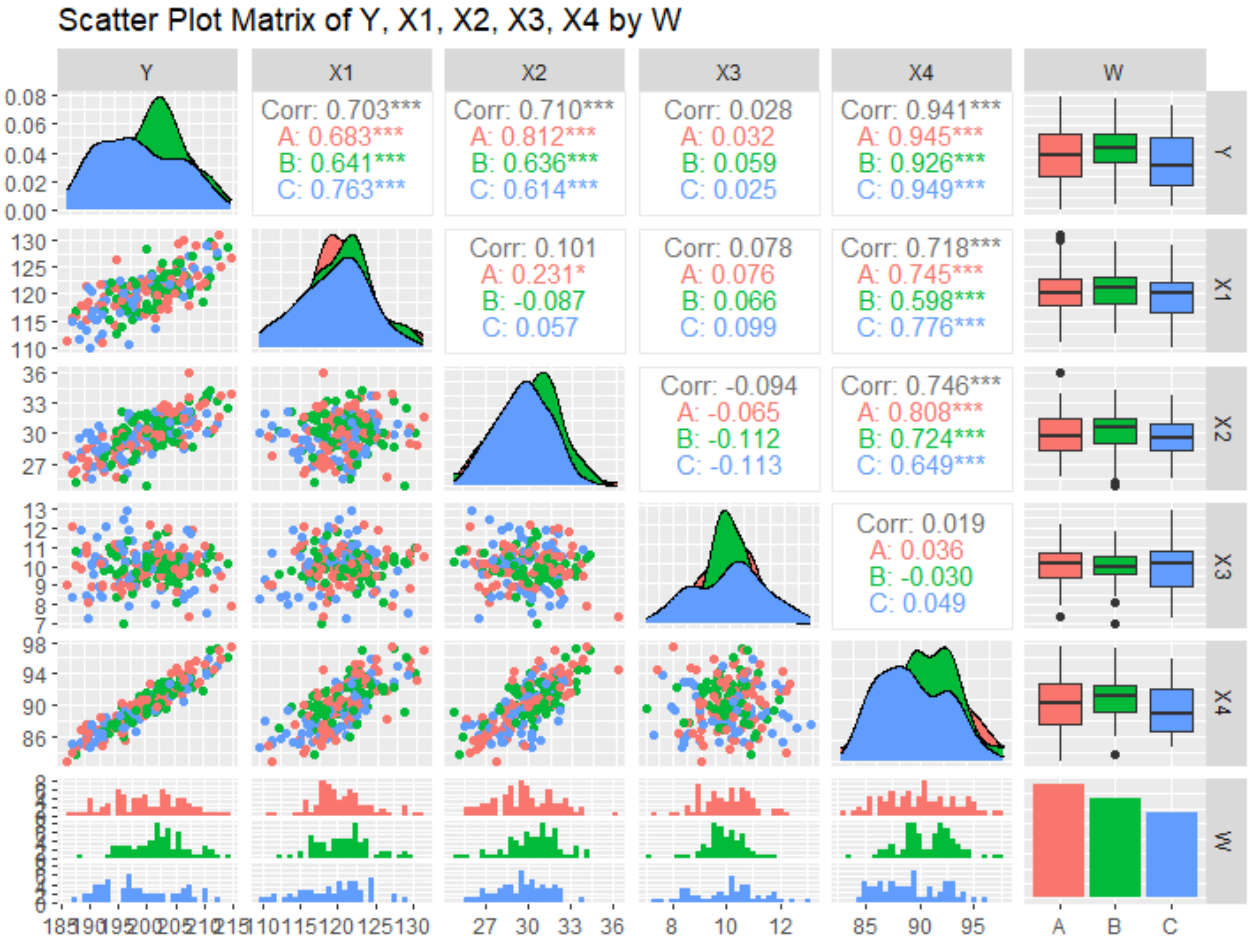




Conclusions:

- Linearity: There is some structure in the residuals vs fitted plot, suggesting that the linearity assumption might be violated.
- Homoscedasticity: There is a clear indication of heteroscedasticity (non-constant variance) in the Scale-Location plot.
- Normality: The residuals seem to follow a normal distribution fairly well, with some minor deviations at the extremes. Therefore, the normality assumption appears to be satisfied.

(b) Provide a scatter-plot matrix of Y, X1, X2, X3, and X4, annotating the different levels of W in each plot using a different color.



- Summary of Insights:
  - Strong Correlations:
    - Y and X4 have a very strong positive correlation (0.941), especially in groups B and C.
    - X2 and X4 also show strong correlations across all groups.
  - Weak or Negative Correlations:
    - X3 shows weak correlations with the other variables, and its correlations with the variables are near zero.
  - Distribution and Group Differences:
    - The distributions of Y, X1, X2, and X4 vary across the levels of W. For instance, the box plots for X4` suggest some differences in the range of values between the groups.

(c) Run the regression model of Y on X4.

```
Regression model of Y on X4
> model_Y_X4 <- lm(Y ~ X4, data = data)
> summary(model_Y_X4)

Call:
lm(formula = Y ~ X4, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5133 -1.3818  0.1039  1.4803  5.9044

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  26.1973     4.4449   5.894  1.6e-08 ***
X4            1.9347     0.0493  39.243  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.129 on 198 degrees of freedom
Multiple R-squared:  0.8861,    Adjusted R-squared:  0.8855
F-statistic:  1540 on 1 and 198 DF,  p-value: < 2.2e-16
```

- The regression model shows a strong relationship between **Y** and **X4**, with residuals ranging from -5.5133 to 5.9044, and a median close to zero (0.1039). Most residuals fall within -1.3818 and 1.4803, with few outliers.
- The coefficient estimates show that **X4** is a significant predictor of **Y**, with an estimate of 1.9347 for **X4**, indicating a positive relationship. Both coefficients are highly significant with p-values well below 0.05.
- The model explains 88.61% of the variance in **Y** (R-squared = 0.8861). The Adjusted R-squared value of 0.8855 confirms that the model fits well, accounting for the number of predictors. The F-statistic of 1540 with a p-value less than 2.2e-16 further supports the model's significance.
- Overall, the model indicates a strong and significant relationship between **Y** and **X4**, with high R-squared values and low p-values indicating a good fit and predictive power.

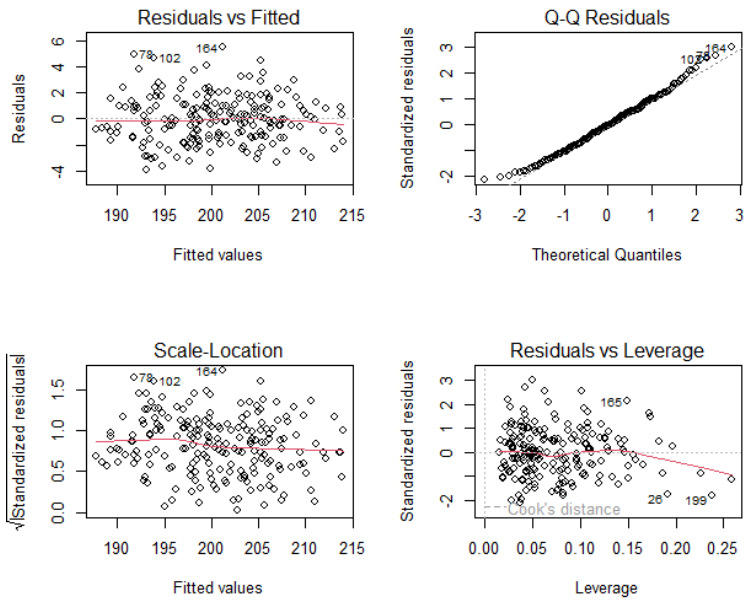
(d) Run the regression model of Y on all the remaining variables (X1, X2, X3, X4, W), including the non-additive terms (i.e., interactions of the continuous predictors with the categorical variable)

```
> # (d) Regression model of Y on all variables, including interactions with W
> cat("\nRegression model of Y on all variables (including interactions with W)\n")

Regression model of Y on all variables (including interactions with W)
> model_Y_all <- lm(Y ~ X1 * W + X2 * W + X3 * W + X4 * W, data = data)
> summary(model_Y_all)

Call:
lm(formula = Y ~ X1 * W + X2 * W + X3 * W + X4 * W, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8807 -1.3656 -0.0337  1.0723  5.4653

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  28.3612     7.1589   3.962 0.000106 ***
X1            1.1682     0.2570   4.545 9.90e-06 ***
WB           -8.2392    11.6561  -0.707 0.480544
WC          -24.4132    10.7774  -2.265 0.024658 *
X2            2.7008     0.5276   5.119 7.64e-07 ***
X3            0.3221     0.2313   1.393 0.165391
X4           -0.5859     0.5015  -1.168 0.244184
X1:WB        -0.2119     0.3432  -0.617 0.537741
X1:WC        -0.4392     0.3618  -1.214 0.226304
WB:X2        -0.9233     0.7186  -1.285 0.200463
WC:X2        -1.3562     0.7368  -1.841 0.067257 .
WB:X3         0.2838     0.3743   0.758 0.449266
WC:X3        -0.3090     0.3076  -1.005 0.316355
WB:X4         0.6572     0.6797   0.967 0.334848
WC:X4         1.3478     0.7030   1.917 0.056730 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.879 on 185 degrees of freedom
Multiple R-squared:  0.9171,    Adjusted R-squared:  0.9108
F-statistic: 146.2 on 14 and 185 DF,  p-value: < 2.2e-16
```

- The dataset includes variables Y (response) and X1, X2, X3, X4, with W consisting of WB and WC. The model investigates the main effects of these variables and their interactions with W. The residuals range from -3.8807 to 5.4653, with most clustered around zero, indicating a good fit. Coefficients for X1 (1.1682) and X2 (2.7008) are highly significant, suggesting they strongly influence Y. Interaction terms like WC:X2 and WC:X4 show marginal significance, indicating some interactions may affect Y.

- The model explains 91.71% of the variance in Y, with a highly significant F-statistic of 146.2 and a low residual standard error of 1.879. WB and WC influence Y, with WC having a significant negative effect (-24.4132). Other interaction terms, like WB:X2 and X1:WC, are not significant, suggesting they do not strongly contribute to the model. The results suggest that WC interacts with other variables to influence Y more than WB.

(e) Examine the regression assumptions and provide alternatives if any of them fail.




```
> # (e) Examine the regression assumptions
> cat("\nExamining regression assumptions\n")

Examining regression assumptions
> par(mfrow = c(2, 2))
> plot(model_Y_all)  # Residuals plots for checking assumptions
>
> # Normality of residuals test using Shapiro-Wilk
> shapiro_test <- shapiro.test(residuals(model_Y_all))
> cat("\nShapiro-Wilk normality test for residuals:\n")

Shapiro-Wilk normality test for residuals:
> shapiro_test

        Shapiro-Wilk normality test

data:  residuals(model_Y_all)
W = 0.9907, p-value = 0.2253
```

Assumption Plots:

- The plots display the following:

Residuals vs Fitted: Shows no clear pattern, indicating a linear relationship.

- Normal Q-Q: Residuals closely follow the normal line.
- Scale-Location: No major deviations from homoscedasticity.
- Residuals vs Leverage: No points seem to exhibit high leverage.
- Overall, the model appears to meet the regression assumptions reasonably well.

(f) Use the "stepwise regression" approach to examine whether you can reduce the dimension of the model.

```
> # (f) Stepwise regression approach using AIC
> cat("\nStepwise regression approach using AIC\n")

Stepwise regression approach using AIC
> stepwise_model <- step(model_Y_all)
Start:  AIC=266.69
Y ~ X1 * W + X2 * W + X3 * W + X4 * W

       Df Sum of Sq    RSS    AIC
- X1:W  2    5.2069 658.33 264.28
- W:X3  2   10.3202 663.44 265.83
- W:X2  2   12.4535 665.58 266.47
- W:X4  2   12.9877 666.11 266.63
<none>              653.12 266.69

Step:  AIC=264.28
Y ~ X1 + W + X2 + X3 + X4 + W:X2 + W:X3 + W:X4

       Df Sum of Sq    RSS    AIC
- W:X3  2     8.731 667.06 262.91
<none>              658.33 264.28
- W:X2  2    20.832 679.16 266.51
- W:X4  2    37.618 695.95 271.39
- X1    1   159.098 817.43 305.57

Step:  AIC=262.91
Y ~ X1 + W + X2 + X3 + X4 + W:X2 + W:X4

       Df Sum of Sq    RSS    AIC
<none>              667.06 262.91
- X3    1    11.587 678.65 264.36
- W:X2  2    21.134 688.20 265.15
- W:X4  2    35.695 702.76 269.34
- X1    1   162.414 829.48 304.50
```

```
> summary(stepwise_model)

Call:
lm(formula = Y ~ X1 + W + X2 + X3 + X4 + W:X2 + W:X4, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0269 -1.2964  0.0009  1.1942  5.6151

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  30.5231     6.7564   4.518 1.10e-05 ***
X1            0.9587     0.1413   6.784 1.46e-10 ***
WB           -7.0505    10.8716  -0.649  0.51743
WC          -29.9678    10.0516  -2.981  0.00325 **
X2            2.2899     0.3218   7.117 2.23e-11 ***
X3            0.2439     0.1346   1.812  0.07159 .
X4           -0.1849     0.2903  -0.637  0.52487
WB:X2        -0.5349     0.2384  -2.244  0.02602 *
WC:X2        -0.4785     0.2481  -1.928  0.05531 .
WB:X4         0.2633     0.1687   1.560  0.12036
WC:X4         0.4966     0.1563   3.178  0.00173 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.879 on 189 degrees of freedom
Multiple R-squared:  0.9153,    Adjusted R-squared:  0.9109
F-statistic: 204.4 on 10 and 189 DF,  p-value: < 2.2e-16
```

Stepwise Regression Model Summary:

The stepwise regression approach using **AIC** has reduced the number of variables in the model based on model performance.

Final Model:

The final model is: $Y = \beta_0 + \beta_1 X_1 + \beta_2 W + \beta_3 X_2 + \beta_4 X_3 + \beta_5 X_4 + \beta_6 (W \times X_2) + \beta_7 (W \times X_4)$

Where:

- **W** is a categorical variable with levels "A", "B", and "C". "WB" and "WC" represent the interaction terms with levels "B" and "C" respectively.
- **Significant Variables**:
  - **X1**: Significant with p-value < 0.001.
  - **WC**: Significant with p-value = 0.003.
  - **X2**: Significant with p-value < 0.001.
  - **W:B:X2**: Significant with p-value = 0.026.
  - **WC:X4**: Significant with p-value = 0.0017.

Model Summary:

- **Multiple R-squared**: 0.9153, indicating that about **91.5%** of the variance in Y is explained by the model.
- **Adjusted R-squared**: 0.9109, showing a slight reduction from the multiple R-squared due to the inclusion of interaction terms.
- **F-statistic**: 204.4, with a p-value < 2.2e-16, confirming that the model is statistically significant.
- **Residual Standard Error**: 1.879, representing the standard deviation of the residuals.

Interpretation of Results:

The stepwise regression reduced the complexity of the model by removing some interaction terms that were not statistically significant, such as those involving X1 and X3. The remaining significant variables help explain the variation in Y more effectively, and the model has an adjusted R-squared of 0.9109, which is still very high.

(g) Using the model found in (f), provide a point estimate and a 95% confidence interval for the prediction of Y when: (X1, X2, X3, X4, W) = (120, 30, 10, 90, B).

```
> # (g) Point estimate and 95% confidence interval for the prediction of Y when:
> # (X1, X2, X3, X4, W) = (120, 30, 10, 90, B)
> cat("\nPoint estimate and 95% confidence interval for the prediction of Y\n")

Point estimate and 95% confidence interval for the prediction of Y
> new_data <- data.frame(X1 = 120, X2 = 30, X3 = 10, X4 = 90, W = "B")
> predict(stepwise_model, new_data, interval = "confidence")
       fit      lwr      upr
1 200.6604 200.1839 201.1369
>
```

The point estimate and the 95% confidence interval for the prediction of Y when the values of (X1, X2, X3, X4, W) are (120, 30, 10, 90, B) are as follows:

- **Point estimate** (fit): 200.6604
- **Lower bound** (lwr): 200.1839
- **Upper bound** (upr): 201.1369

This means that, with 95% confidence, the value of Y is expected to fall between 200.1839 and 201.1369 when the predictor values are set to (120, 30, 10, 90, B).

(h) Using the cut() function, create a categorical variable (named Z) with 3 levels based on the quantiles of X4. Provide the contingency table of Z 3 and W

```
> # (h) Create a categorical variable z based on the quantiles of X4, and provide
the contingency table
> cat("\nCreating a categorical variable Z based on the quantiles of X4\n")

Creating a categorical variable Z based on the quantiles of X4
> data$Z <- cut(data$X4, breaks = quantile(data$X4, probs = 0:3 / 3), labels = c
("Low", "Medium", "High"))
> cat("\nContingency table of Z and W:\n")

Contingency table of Z and W:
> table(data$Z, data$W)

          A  B  C
  Low    26 13 27
  Medium 24 27 15
  High   25 27 15
>
```

Contingency Table Interpretation:

- **Row "Low":**

26 values of X4 fall into the "Low" category and correspond to W = A.

13 values of X4 fall into the "Low" category and correspond to W = B.

27 values of X4 fall into the "Low" category and correspond to W = C.

- **Row "Medium":**

24 values of X4 fall into the "Medium" category and correspond to W = A.

27 values of X4 fall into the "Medium" category and correspond to W = B.

15 values of X4 fall into the "Medium" category and correspond to W = C.

- **Row "High":**

25 values of X4 fall into the "High" category and correspond to W = A.

27 values of X4 fall into the "High" category and correspond to W = B.

15 values of X4 fall into the "High" category and correspond to W = C.

Key Insights:

The distribution of Z (Low, Medium, High) across the levels of W is fairly consistent.

For each level of W, the values of X4 are relatively evenly distributed across the "Low", "Medium", and "High" categories, with a slight emphasis on the "High" category for most levels of W.

(i) Run the parametric two-way ANOVA of Y on the categorical variables W and Z (including the interaction term). Provide the fit, examine the assumptions, and comment on the significance of the terms.

```
> # (i) Run the parametric two-way ANOVA of Y on W and Z, including the interaction
term
> cat("\nRunning two-way ANOVA of Y on W and Z, including the interaction term\n")

Running two-way ANOVA of Y on W and Z, including the interaction term
> anova_two_way <- aov(Y ~ W * Z, data = data)
> summary(anova_two_way)
             Df Sum Sq Mean Sq F value   Pr(>F)
W             2    328   164.2  19.102 2.76e-08 ***
Z             2   5704  2852.0 331.883  < 2e-16 ***
W:Z           4     28     6.9   0.808    0.521
Residuals   190   1633     8.6
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
1 observation deleted due to missingness
>
> # Check the assumptions of the two-way ANOVA
> cat("\nChecking assumptions of two-way ANOVA\n")

Checking assumptions of two-way ANOVA
> par(mfrow = c(2, 2))
> plot(anova_two_way)
```

Diagnostic Plots Interpretation:

1. **Residuals vs Fitted**: Residuals are randomly scattered around 0, indicating a good model fit.
2. **Q-Q Plot**: Points follow the diagonal, suggesting residuals are approximately normally distributed.
3. **Scale-Location**: Residuals are evenly spread, supporting the assumption of constant variance (homoscedasticity).
4. **Residuals vs Leverage**: Some points exceed the Cook's distance threshold, indicating potential influential points.

ANOVA Results:

1. **W (Factor)**: Significant effect (p = 2.76e-08), explaining substantial variance.
2. **Z (Factor)**: Highly significant (p < 2e-16), with a large effect.
3. **W:Z (Interaction)**: No significant interaction (p = 0.521).
4. **Residuals**: Some unexplained variance, but model fits well overall.

Summary:

- **W and Z** are significant predictors.
- **No significant interaction** between W and Z.
- Model assumptions are mostly met, with some potential influential points.