Academic Semester: 2023-2024

Probability and Statistics for Data Analysis

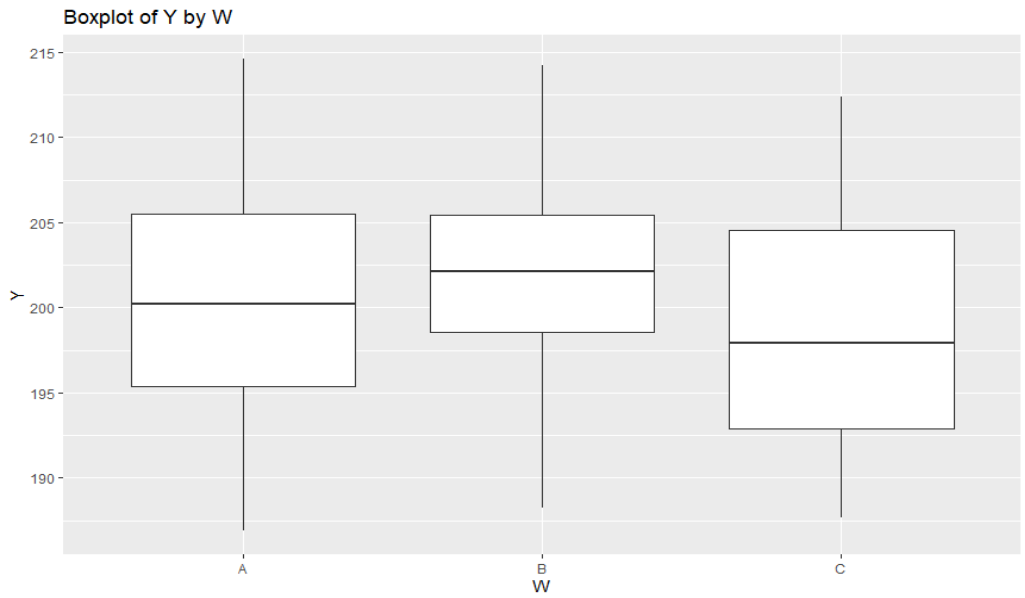
Assignment 2 report

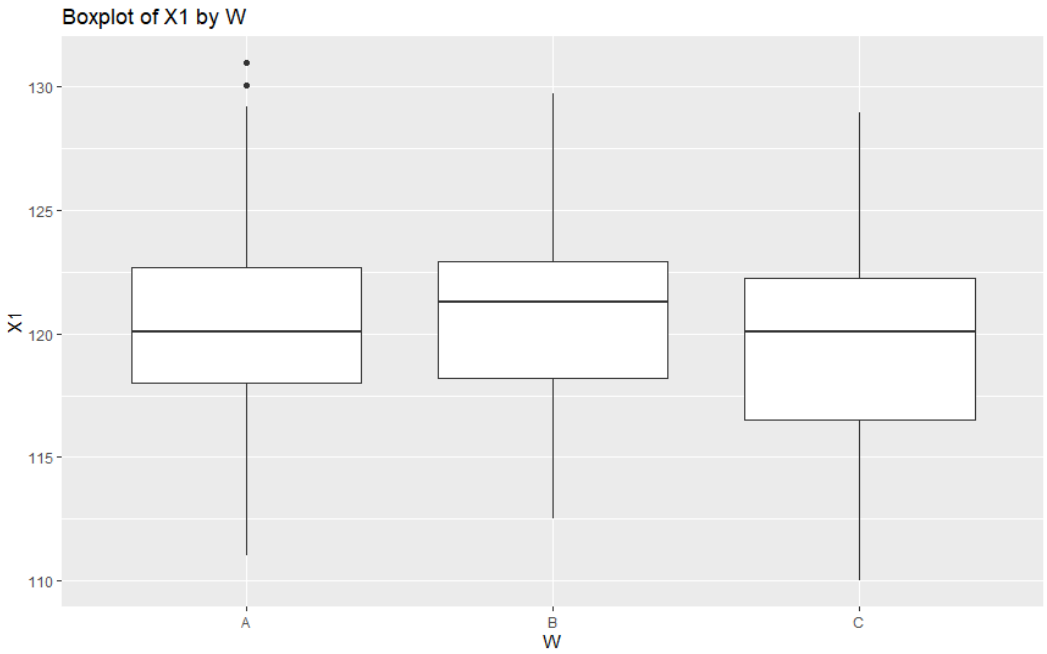
Full name: Vasileios Ilias Drouzas

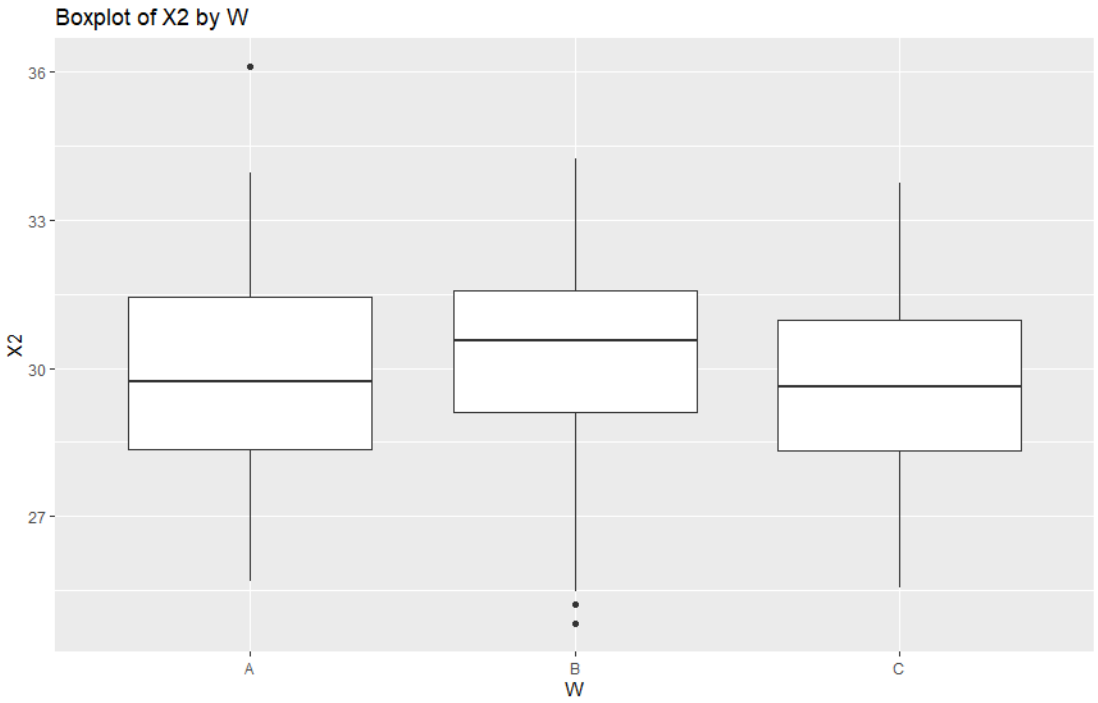
ID: f3352301

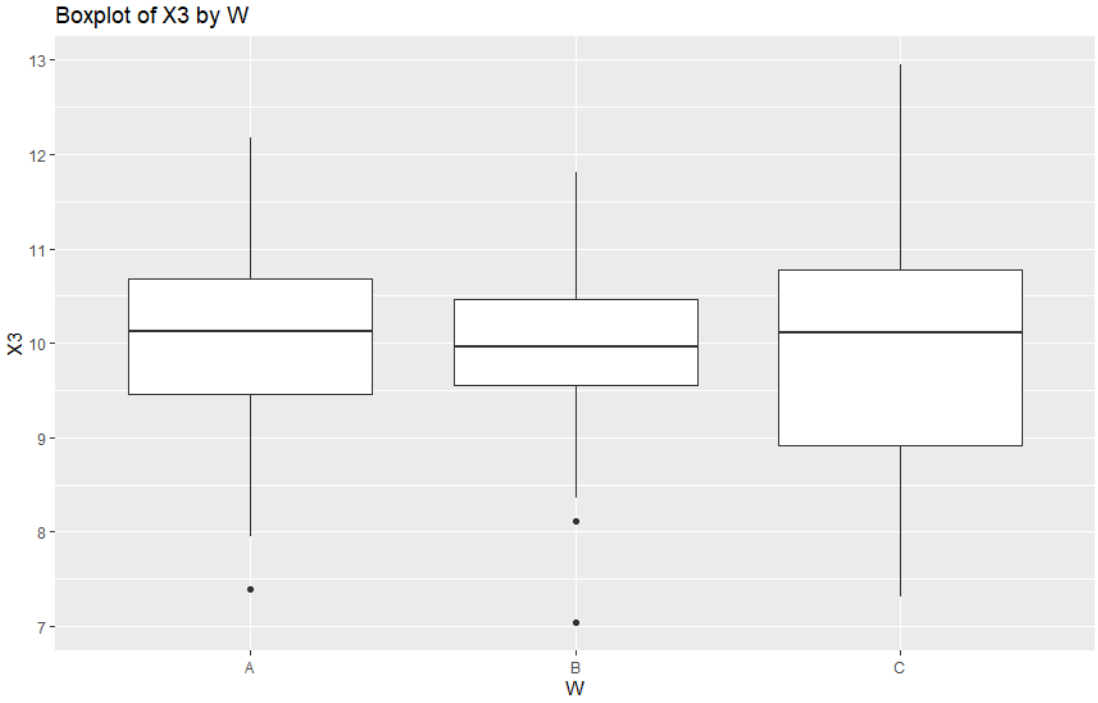
Exercise 1

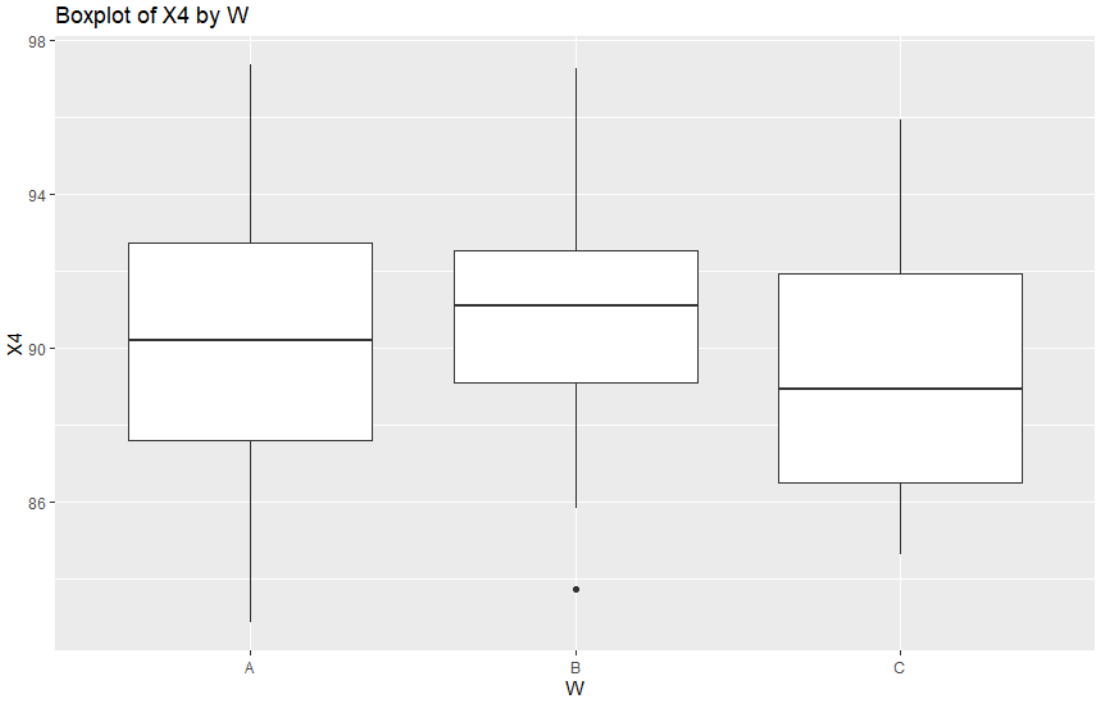
a. (i) We will provide the boxplots of the continuous variables on the categorical variable W:






Boxplot of X2 by W
36
33
30
27
X2
A
B
C
W


Boxplot of X3 by W
13
12
11
10
9
8
7
X3
A
B
C
W



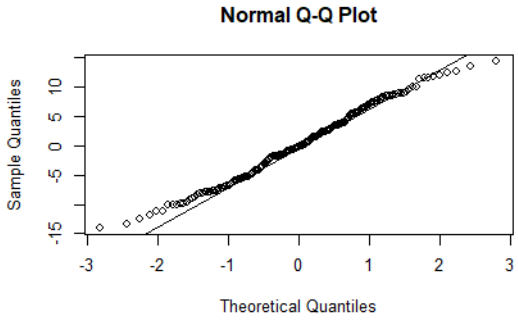
(ii) ANOVA outputs:

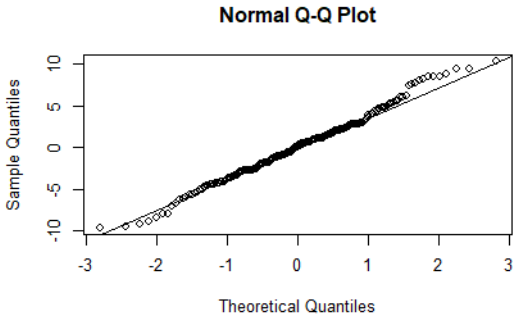
```
> summary(anova_Y)
             Df Sum Sq Mean Sq F value Pr(>F)
W             2    333  166.71   4.352 0.0141 *
Residuals   197   7546   38.31
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(anova_X1)
             Df Sum Sq Mean Sq F value Pr(>F)
W             2   76.3   38.13    2.42 0.0915 .
Residuals   197 3104.1   15.76
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> 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
> 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
> 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
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

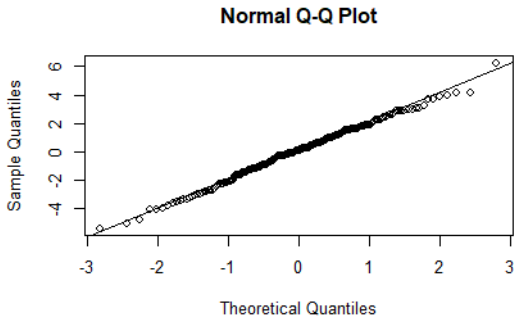
(iii) Checking the assumptions.

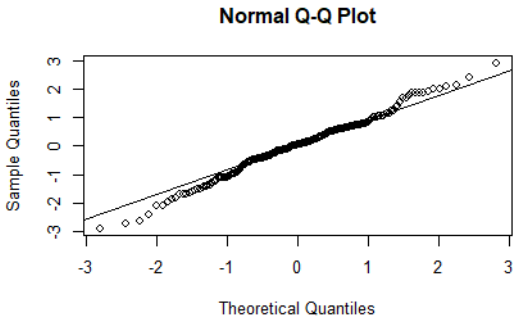
1. Normality of residuals.

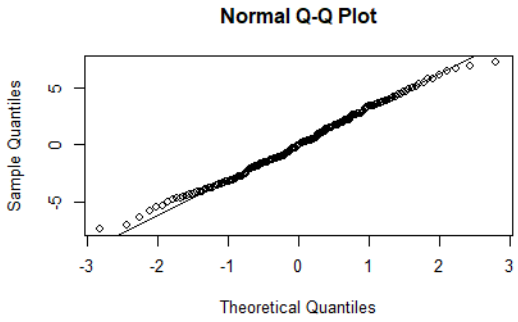
(in order: Y, X1,X2,X3,X4)











From the first look, the residuals seem to follow normality. Let's make ourselves more confident with the Shapiro Wilk test:

```
> shapiro.test(residuals_Y)

        Shapiro-Wilk normality test

data:  residuals_Y
W = 0.98923, p-value = 0.1374

> shapiro.test(residuals_X1)

        Shapiro-Wilk normality test

data:  residuals_X1
W = 0.99123, p-value = 0.268

> shapiro.test(residuals_X2)

        Shapiro-Wilk normality test

data:  residuals_X2
W = 0.99539, p-value = 0.8049

> shapiro.test(residuals_X3)

        Shapiro-Wilk normality test

data:  residuals_X3
W = 0.99108, p-value = 0.2555

> shapiro.test(residuals_X4)

        Shapiro-Wilk normality test

data:  residuals_X4
W = 0.99272, p-value = 0.4243
```

Since the p-value is above 0.05 for all occasions on 95% confidence level, we cannot reject normality for each case.

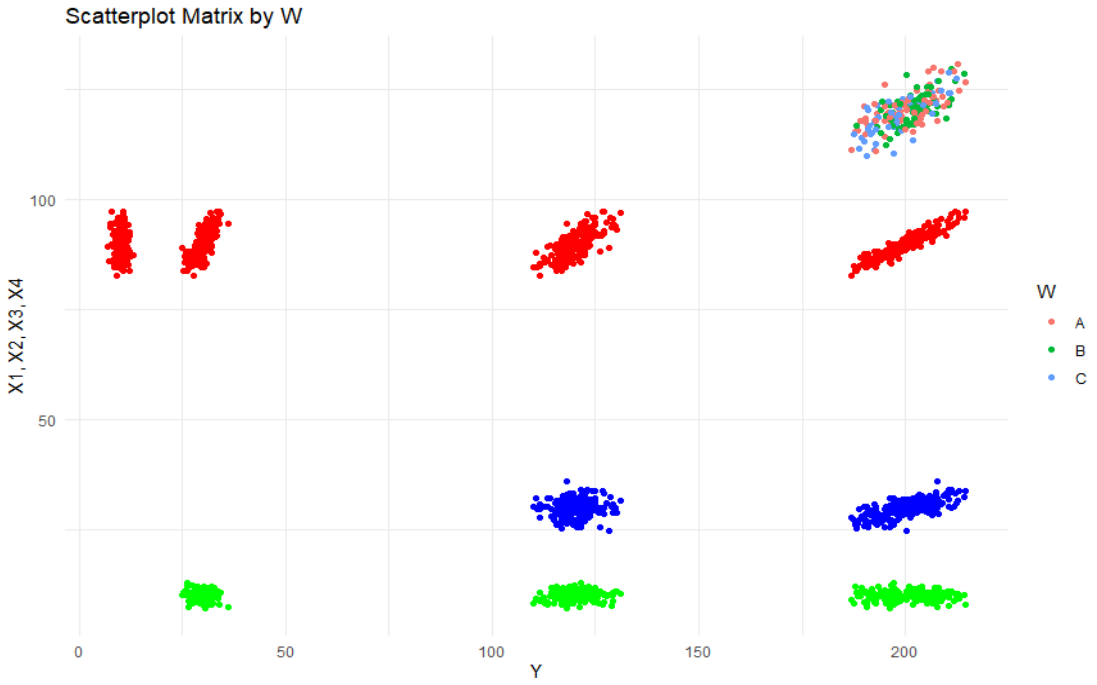
2. Homogeinity of variances:

We will use Levene's test here:

```
> #Levene's test
> leveneTest(anova_Y)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value  Pr(>F)
group   2  3.6897 0.02672 *
      197
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> leveneTest(anova_X1)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   2  1.0945 0.3367
      197
> leveneTest(anova_X2)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   2   1.081 0.3412
      197
> leveneTest(anova_X3)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value    Pr(>F)
group   2  7.4498 0.0007605 ***
      197
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> leveneTest(anova_X4)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   2  2.2203 0.1113
      197
> |
```

Here, the ANOVA models for Y vs W and X3 vs W do not hold the assumption of homogeinity of variances.

b. The scatterplot of Y X1,X2,X3 and X4 is the following:



c. The regression model is the following:

```
> model_x4 <- lm(Y ~ X4, data = data)
> summary(model_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
```

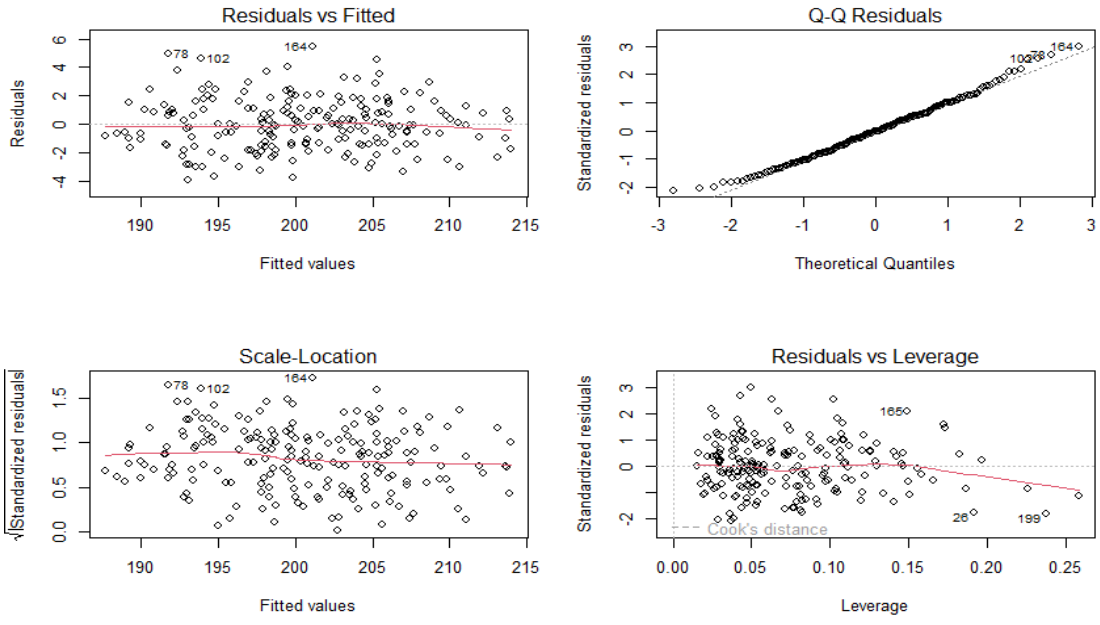
The model seems to be statistically significant as indicated by the extremely low p-value (< 2.2e-16). In general, it seems to be a good fit for predicting Y based on X4, given

the low p-values, high R-squared, and overall statistical significance.

d. The new model is:

```
> model_all <- lm(Y ~ X1 * W + X2 * W + X3 * W + X4 * W, data = data)
> summary(model_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
```

e. Let's check the assumptions of the last model:



Residuals vs fitted: Horizontal line without distinct patterns, indicating absence of non-linear models.

Normal-QQ: The residuals follow a straight line, indicating normality.

Scale-Location: Interpreting similarly to the first case, homoscedasticity is secured.

Residuals vs Leverage: We watch out for outlying values at the upper right corner or at the lower right corner. Those spots are the places where cases can be influential against a regression line. There is no influential case, or cases.

All in all, I consider the assumptions met in this case.

f. Let's implement the stepwise seletion procedure:

```
> final_model <- step(model_all, direction = "both")
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
+ X1:W  2     5.207 653.12 266.69
- 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
+ W:X3  2     8.731 658.33 264.28
- X3    1    11.587 678.65 264.36
- W:X2  2    21.134 688.20 265.15
+ X1:W  2     3.618 663.44 265.83
- W:X4  2    35.695 702.76 269.34
- X1    1   162.414 829.48 304.50
```

- The stepwise procedure sequentially removes terms that do not contribute significantly to reducing the AIC, aiming to find the best-fitting model based on this criterion.
- The final selected model, with an AIC of 262.91, includes X1, W, X2, X3, X4, and interactions W:X2 and W:X4. These variables are considered to provide the best balance between model complexity and goodness of fit according to the AIC.

  Summarizing the final model, we have:

```
> summary(final_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
```

g.

- Point estimate: 200.6604
- 95% confidence interval: 200.1839 201.1369

h. The contingency table of X4 and W is the following:

	A	B	C
82.86	1	0	0
83.74	0	1	0
83.8	1	0	0
84.42	1	0	0
84.63	0	0	1
84.74	1	0	0
84.83	1	0	0
84.95	0	0	1
84.98	0	0	1
85.16	0	0	1
85.18	0	0	1
85.36	0	0	1
85.39	1	0	0
85.45	0	0	1
85.47	1	0	0
85.61	1	0	0
85.68	1	0	0
85.73	0	0	1
85.81	0	0	1
85.84	0	1	0
86.06	0	0	1
86.13	0	0	1
86.31	0	0	1
86.33	1	0	0
86.36	0	0	1
86.4	1	0	0
86.44	0	0	1
86.45	0	1	0
86.51	0	0	1
86.58	1	0	0
86.73	0	1	0
86.75	1	0	0
86.79	1	0	0

86.93 1 0 0
87.01 0 1 0
87.09 1 0 0
87.12 0 0 1
87.2 0 1 0
87.23 0 0 1
87.43 1 0 0
87.44 1 0 0
87.46 0 0 1
87.5 0 1 0
87.52 1 0 0
87.59 0 0 1
87.6 1 0 1
87.63 0 1 0
87.68 1 0 0
87.71 0 1 0
87.74 0 0 1
87.8 0 0 1
87.81 0 1 0
87.87 1 0 0
87.97 0 0 1
88.19 0 0 1
88.23 1 0 0
88.26 0 1 0
88.3 0 0 1
88.35 1 0 0
88.58 0 0 1
88.62 0 0 1
88.63 1 0 0
88.65 0 1 0
88.66 1 0 0
88.67 1 0 0
88.76 0 1 0
88.84 0 1 0
88.9 0 0 1
88.91 1 0 0
88.92 0 0 1

88.93 0 1 0
88.96 0 1 0
88.99 0 1 0
89 1 0 0
89.04 1 0 0
89.14 1 0 0
89.16 0 0 1
89.18 0 1 0
89.24 1 0 0
89.25 0 1 0
89.31 0 0 1
89.32 0 0 1
89.35 0 0 1
89.36 0 1 0
89.37 0 0 1
89.45 1 0 0
89.49 0 0 1
89.5 0 2 0
89.57 0 1 0
89.58 0 1 0
89.59 0 0 1
89.66 0 1 0
89.74 1 1 0
89.76 0 1 0
89.86 1 0 0
89.88 0 0 1
89.89 0 1 0
89.9 1 0 0
89.91 1 0 0
89.95 0 1 0
90.06 0 1 0
90.14 1 0 0
90.16 0 1 0
90.27 1 0 0
90.28 0 1 0
90.43 0 0 1
90.45 1 0 0

90.48 1 0 0
90.53 1 0 0
90.55 1 0 0
90.57 1 0 0
90.64 1 0 0
90.72 0 1 0
90.77 1 0 0
91.1 0 1 0
91.12 0 1 0
91.15 1 0 0
91.18 1 0 0
91.22 0 0 2
91.23 0 0 1
91.26 0 1 0
91.3 1 0 0
91.33 0 1 0
91.36 0 1 0
91.47 0 1 0
91.57 1 0 0
91.69 1 0 0
91.71 0 1 1
91.78 0 1 0
91.81 1 0 0
91.85 0 2 0
91.91 2 0 0
91.93 0 0 1
91.96 0 0 1
92.17 0 1 0
92.2 1 0 0
92.21 0 1 0
92.25 0 2 0
92.27 1 0 0
92.29 0 1 0
92.36 0 1 0
92.47 0 0 1
92.48 0 1 0
92.57 0 0 1

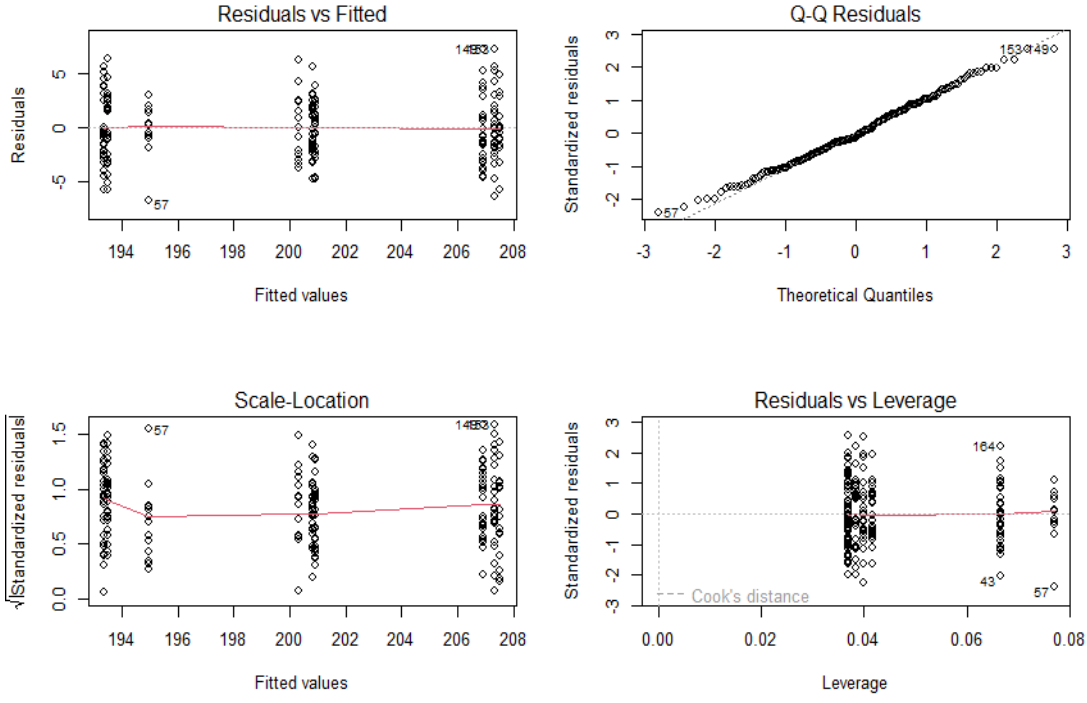
92.58 0 1 0
92.61 0 1 0
92.62 0 1 0
92.66 0 0 1
92.72 1 0 0
92.75 0 0 1
92.78 1 0 0
92.79 0 0 1
92.87 1 0 0
92.9 0 0 1
92.92 0 1 0
92.93 1 0 1
92.95 0 1 0
92.96 0 1 0
92.99 0 0 1
93.01 0 1 0
93.16 1 0 0
93.34 0 0 1
93.39 0 1 0
93.44 0 1 0
93.46 1 0 0
93.52 1 0 0
93.65 0 1 0
93.85 0 0 1
94.02 1 0 0
94.13 0 0 1
94.15 1 0 0
94.16 0 1 0
94.23 0 1 0
94.29 1 0 0
94.33 1 0 0
94.42 0 1 0
94.54 1 0 0
94.56 1 0 0
94.62 0 1 0
94.78 1 0 0
94.98 0 0 1

95.17 1 0 0
95.18 1 0 0
95.6 1 0 0
95.93 1 0 1
96.11 0 1 0
96.91 0 1 0
97.06 1 0 0
97.26 0 1 0
97.36 1 0 0

i. The two way ANOVA of Y on W and Z model.

```
> model <- aov(Y ~ W * Z, data = data)
> summary(model)
             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
```

Both factors W and Z individually have a significant impact on the dependent variable, but their interaction does not contribute significantly to explaining the variation in the dependent variable. Let’s check the assumptions:



Residuals vs fitted: Here, we see that clearly 3 teams(patterns) are formed, indicating non-linearity.

Normal-QQ: The residuals follow a straight line, indicating normality.

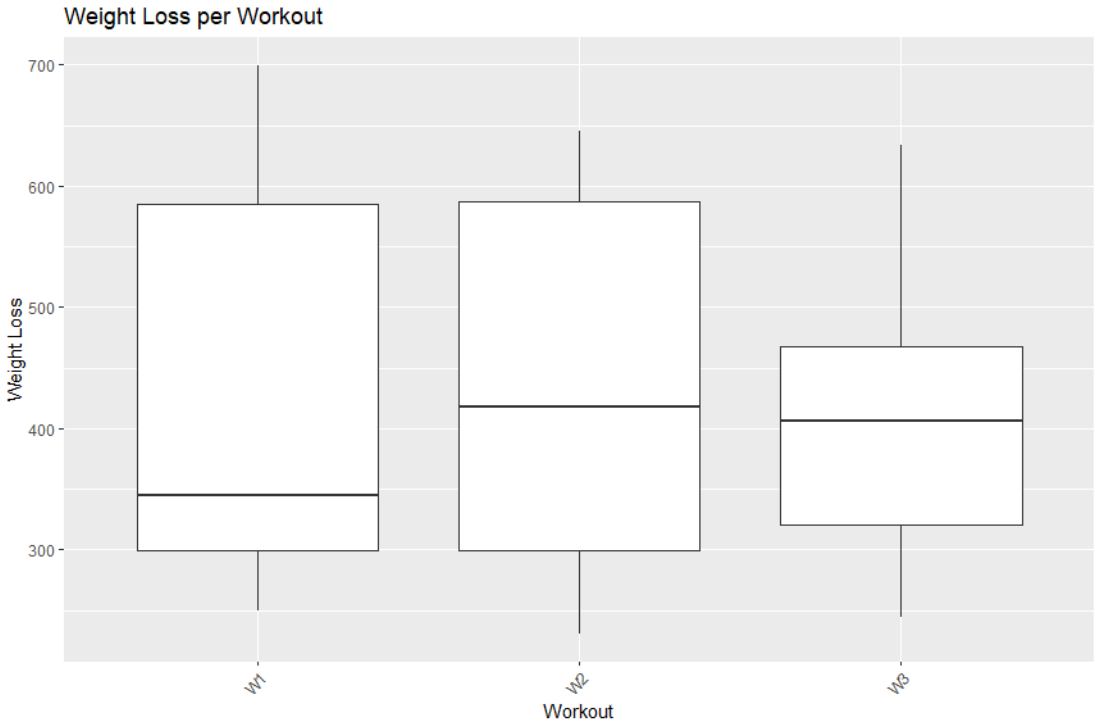
Scale-Location: Interpreting similarly to the first case, homoscedasticity is not secured.

Residuals vs Leverage: We watch out for outlying values at the upper right corner or at the lower right corner. Those spots are the places where cases can be influential against a regression line. We cannot be sure about it here.

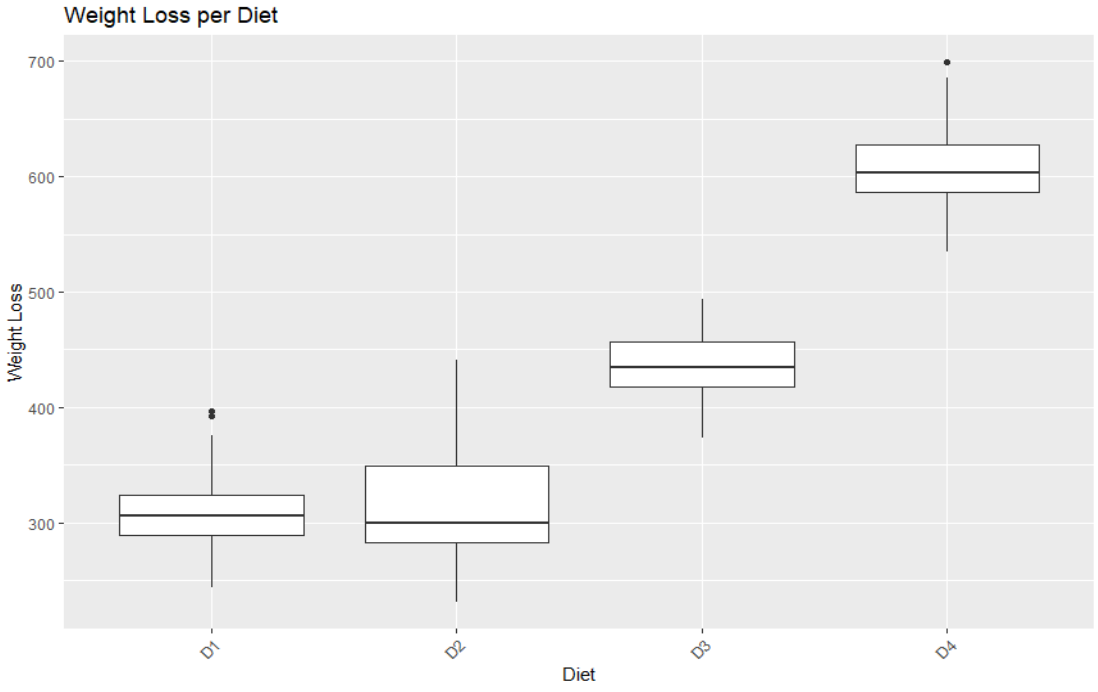
Clearly, the assumptions are not satisfied, so the model we fitted here is not adequate for further investigation and predictions.

Exercise 2

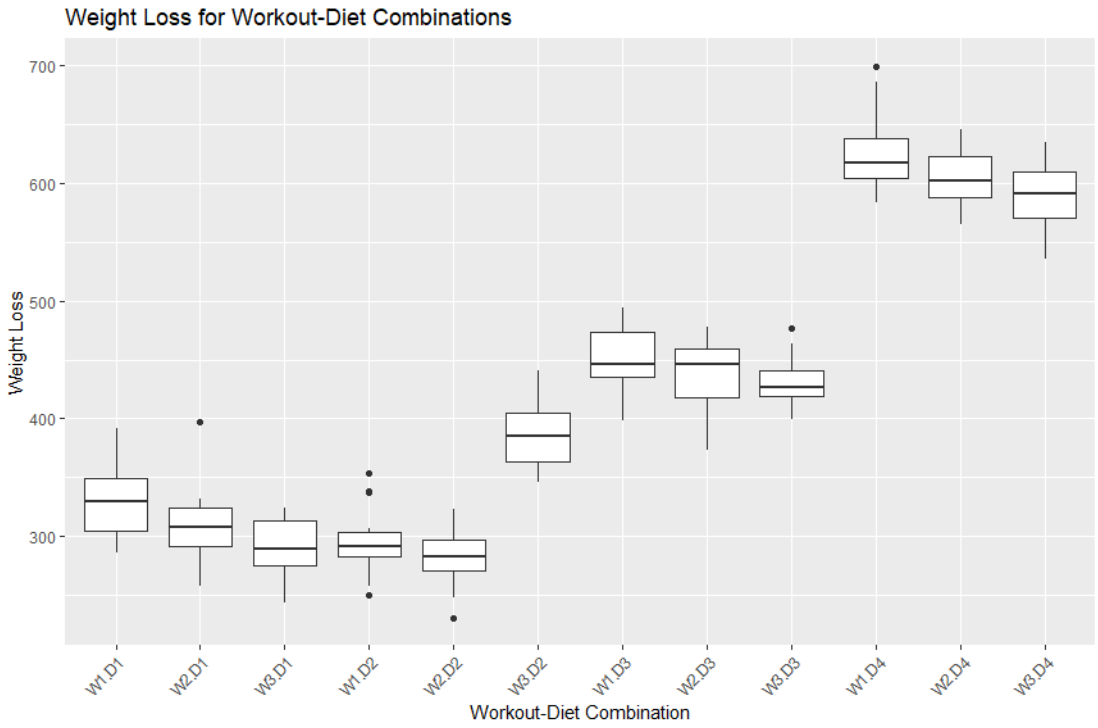
(a)Boxplot of weight loss per workout:



Boxplot of weight loss per diet:



Boxplot of weight loss for combinations of workout and diet:



b. Fitting one-way ANOVA.

```
> model <- aov(data$loss ~ data$workout, data = data)
> #summary of the ANOVA model
> summary(model)
              Df  Sum Sq Mean Sq F value Pr(>F)
data$workout   2   13831    6916     0.4  0.671
Residuals    237 4101316   17305
```

Model parameters:

- Df: Represents the degrees of freedom (2 for the workout variable, 237 for the residuals).
- Sum Sq: Sum of squares represents the variability in weight loss which is explained by the workout factor.
- Mean Sq: Reflects the average variability among the groups.
- F value: It represents whether there are significant differences between the weight loss among the workout groups. Here, it has a value of 0.4.

- Pr (p-value): A p-value of 0.671 suggests weak evidence against the null hypothesis and implies that there isn't enough statistical evidence to reject the idea that all workout groups have the same mean weight loss.

Our conclusion is that there are no statistically significant differences in weight loss among the different workout groups.

c. Changing the reference level.

```
> new_model <- aov(data$loss ~ data$workout, data = data)
> summary(new_model)
              Df  Sum Sq Mean Sq F value Pr(>F)
data$workout   2   13831    6916     0.4  0.671
Residuals    237 4101316   17305
```

But let’s also have a t-test between the data loss of W2 and W3:

```
        Welch Two Sample t-test

data:  w2_data$loss and w3_data$loss
t = 0.73796, df = 159.96, p-value = 0.4616
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -24.16899  53.00723
sample estimates:
mean of x mean of y
 430.6941  416.2750
```

Here, we have a big p-value, indicating that we cannot reject that there is no true difference in the means of the two groups. So there is no statistically significant difference between the two groups.

d.

```
> model2 <- aov(data$loss ~ data$diet, data = data)
> summary(model2)
             Df  Sum Sq Mean Sq F value Pr(>F)
data$diet     3 3803266 1267755   959.3 <2e-16 ***
Residuals   236  311882    1322
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model parameters:

- Df: Represents the degrees of freedom (3 for the diet variable, 236 for the residuals).
- Sum Sq: Sum of squares represents the variability in weight loss which is explained by the diet factor.
- Mean Sq: Reflects the average variability among the groups.
- F value: The ratio of the Mean Square for the 'diet' factor to the mean square of the residuals.It measures whether there are significant differences in weight loss among the diet groups.
- Pr (p-value): The p-value (<0.001), suggests that there are significant differences between weight loss and diet groups.

Not all treatments are found to be significant. The treatments that show non-significant differences are D1 and D2.

e. Now let's exclude D1 and D2. Our new model looks like this:

```
> summary(new_model2)
             Df Sum Sq Mean Sq F value Pr(>F)
diet          1 799840  799840     993 <2e-16 ***
Residuals   113  91018     805
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After excluding 'D1' and 'D2', the ANOVA for the 'diet' factor indicates that there are still significant differences in weight loss among the remaining diet types ('D3', 'D4', 'D5').

The extremely low p-value (<2e-16) strongly suggests that at least one of these remaining diet types significantly differs from the others in terms of weight loss.

The high F value further supports this, showing substantial variability among the remaining diet groups.

f. Now, we will fit a two-way ANOVA model of main effects:

```
> #Two-Way ANOVA model
> model <- aov(data$loss ~ data$workout + data$diet, data = data)
> summary(model)
              Df  Sum Sq Mean Sq F value  Pr(>F)
data$workout   2   13831    6916   5.349 0.00535 **
data$diet      3 3798759 1266253 979.329 < 2e-16 ***
Residuals    234  302557    1293
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The extremely low p-value (< 2e-16, which is essentially zero) associated with "diet" suggests strong evidence against the null hypothesis. The p value for workout is also small enough. Both ‘workout’ and diet’ factors have statistically significant effects on weight loss.

g. After some search, we find out that workout does not have insignificant levels and ‘diet’ has D3 as a non-significant level, so we will remove it.

```
> non_significant_workout
character(0)
> non_significant_diet
[1] "D3"
>
```

Filtering D3 out, we refit the model and get the results:

```
> model_simplified <- aov(data$loss ~ data$workout + data$diet, data = data_filtered)
> summary(model_simplified)
              Df  Sum Sq Mean Sq F value  Pr(>F)
data$workout   2   13831    6916   5.349 0.00535 **
data$diet      3 3798759 1266253 979.329 < 2e-16 ***
Residuals    234  302557    1293
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We notice that the results are the same as before. This validates that D3 did not have a significant impact one the overall model.

h. Let's fit a two-way ANOVA model with interactions:

```
> model_interaction <- aov(data$loss ~ data$workout * data$diet, data = data)
> summary(model_interaction)
                       Df  Sum Sq Mean Sq  F value   Pr(>F)
data$workout            2   13831    6916    9.836 7.99e-05 ***
data$diet               3 3798759 1266253 1800.976  < 2e-16 ***
data$workout:data$diet  6  142252   23709   33.721  < 2e-16 ***
Residuals             228  160305     703
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model parameters:

- Df: 2 degrees of freedom for the 'workout' variable, 3 for the 'diet' variable and 6 for the interaction between 'workout' and 'diet'.
- Sum of Squares, mean squares: Similar interpretation as before.
- All f values are high.
- The significant p-values (<< 0.001) for 'workout', 'diet', and their interaction indicate that both factors individually and their interaction significantly influence weight loss.

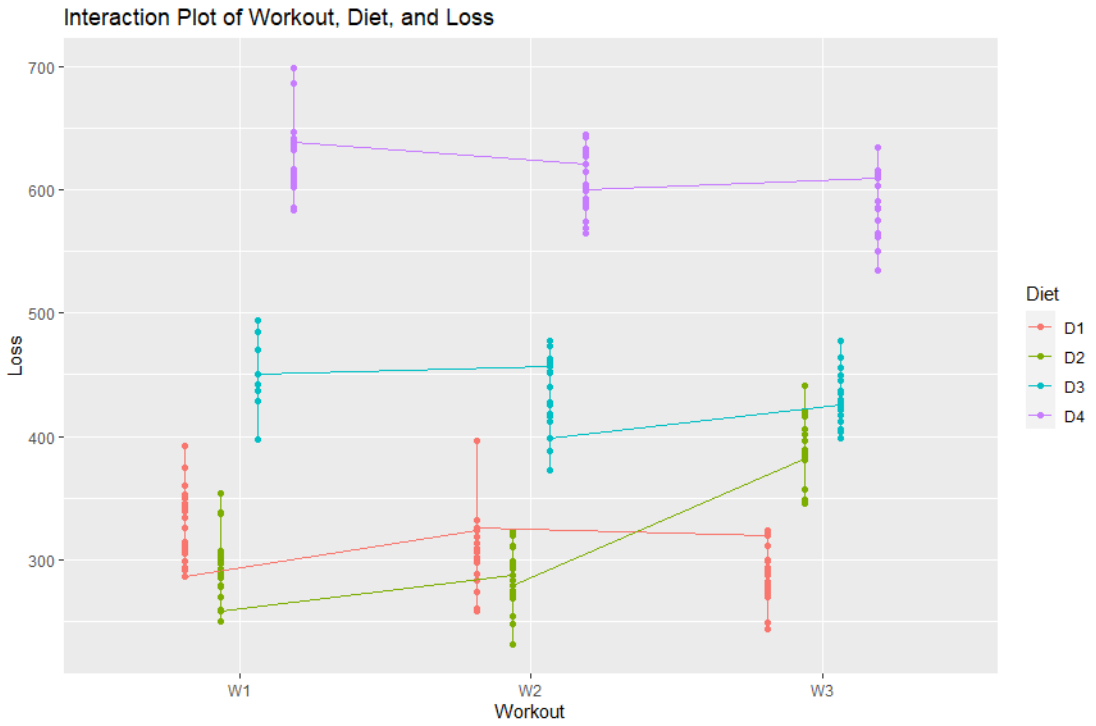
All parameters have very small p-values, way smaller than the significance level, so they can be safely be considered significant.

i. Now we will do stepwise selection:

```
> stepwise_model <- step(model_interaction, direction = "both", trace = FALSE)
> summary(stepwise_model)
                       Df  Sum Sq Mean Sq  F value   Pr(>F)
data$workout            2   13831    6916    9.836 7.99e-05 ***
data$diet               3 3798759 1266253 1800.976  < 2e-16 ***
data$workout:data$diet  6  142252   23709   33.721  < 2e-16 ***
Residuals             228  160305     703
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

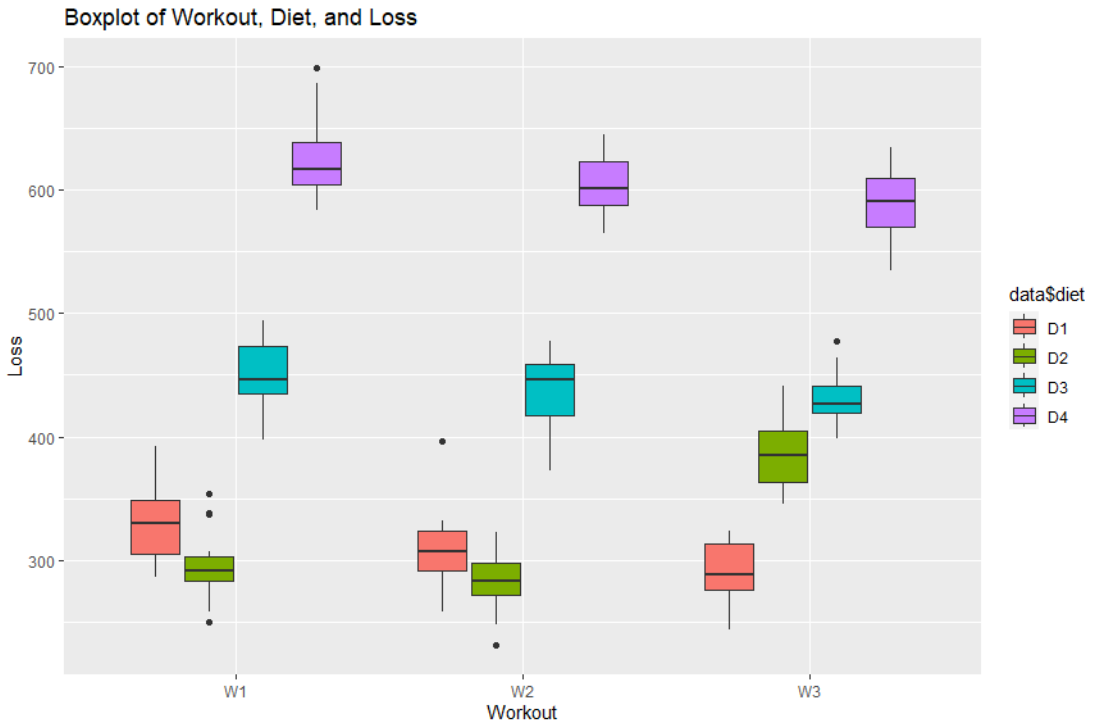
This is the final model, which is chosen from the stepwise selection, minimizing the AIC. It is exactly the same model with the starting model, meaning that the stepwise selection did not suggest any changes to the starting model. Based on the AIC model, this model remains the best. Of course, all parameters are significant here too.

j. Let’s make an interaction model:



This plot shows the interaction between "Workout" and "Loss" while considering the effect of "Diet."

Alternatively, we can make a boxplot of these three:



k. Summarizing the three models, we have:

```
> summary(constant_model)            # constant model
             Df  Sum Sq Mean Sq F value Pr(>F)
Residuals   239 4115147   17218
> summary(model)                     # main effects model
              Df  Sum Sq Mean Sq F value  Pr(>F)
data$workout   2   13831    6916   5.349 0.00535 **
data$diet      3 3798759 1266253 979.329 < 2e-16 ***
Residuals    234  302557    1293
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(model_interaction)         # interaction model
                      Df  Sum Sq Mean Sq  F value   Pr(>F)
data$workout           2   13831    6916    9.836 7.99e-05 ***
data$diet              3 3798759 1266253 1800.976  < 2e-16 ***
data$workout:data$diet   6  142252   23709   33.721  < 2e-16 ***
Residuals            228  160305     703
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The constant model is the simplest model that assumes no predictor variables and uses only the mean to predict the outcome. Its residuals are substantially higher compared to the other models.
- Both the main effects model and the interaction model show significant F-values for workout and diet factors, indicating that these models significantly outperform the constant model.
- The interaction model, however, includes an interaction term (workout:diet) and exhibits additional significantly low p-values for this term. This suggests that the interaction model might provide a better fit compared to the main effects model as it accounts for the interaction effect between workout and diet.

In conclusion, the interaction model appears to provide the best fit among the models considered, as it captures the combined effects and interactions between workout and diet factors more comprehensively compared to the main effects model or the constant model.

Finally, we can run an ANOVA check for the three models:

```
> anova_result <- anova(constant_model, model, model_interaction)
> print(anova_result)
Analysis of Variance Table

Model 1: data$loss ~ 1
Model 2: data$loss ~ data$workout + data$diet
Model 3: data$loss ~ data$workout * data$diet
  Res.Df     RSS Df Sum of Sq        F    Pr(>F)
1    239 4115147
2    234  302557  5   3812590 1084.520 < 2.2e-16 ***
3    228  160305  6    142252   33.721 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Both Model 2 and Model 3 show significant improvements over the previous models, as indicated by their highly significant p-values.
- Model 3 (with the interaction term) appears to provide a better fit compared to Model 2, as it further reduces the RSS.
- The interaction between workout and diet seems to contribute significantly to explaining the variability in the loss variable beyond just the individual effects of workout and diet.

If I were to choose a model among the three for predicting 'loss' based on the variables 'workout' and 'diet', I would choose the third one, which involves the interaction term.