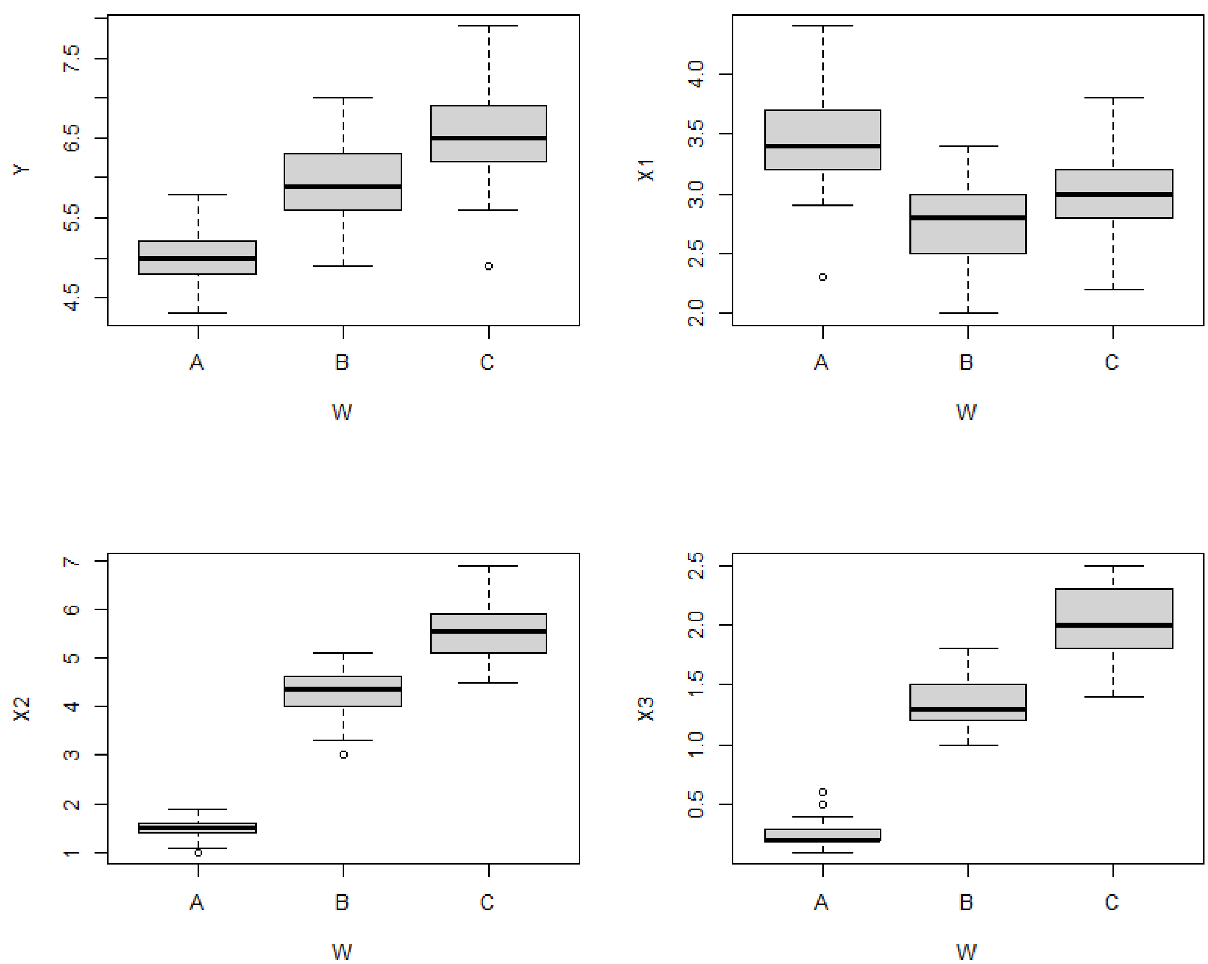
1.

(a)

(i)*To begin with, I preprocessing our given data in order to use them properly, so I manage with the punctuation problem.*
Afterwards, I depict the variables Y, X1, X2, X3 along with categorical variable W with the boxplots bellow:



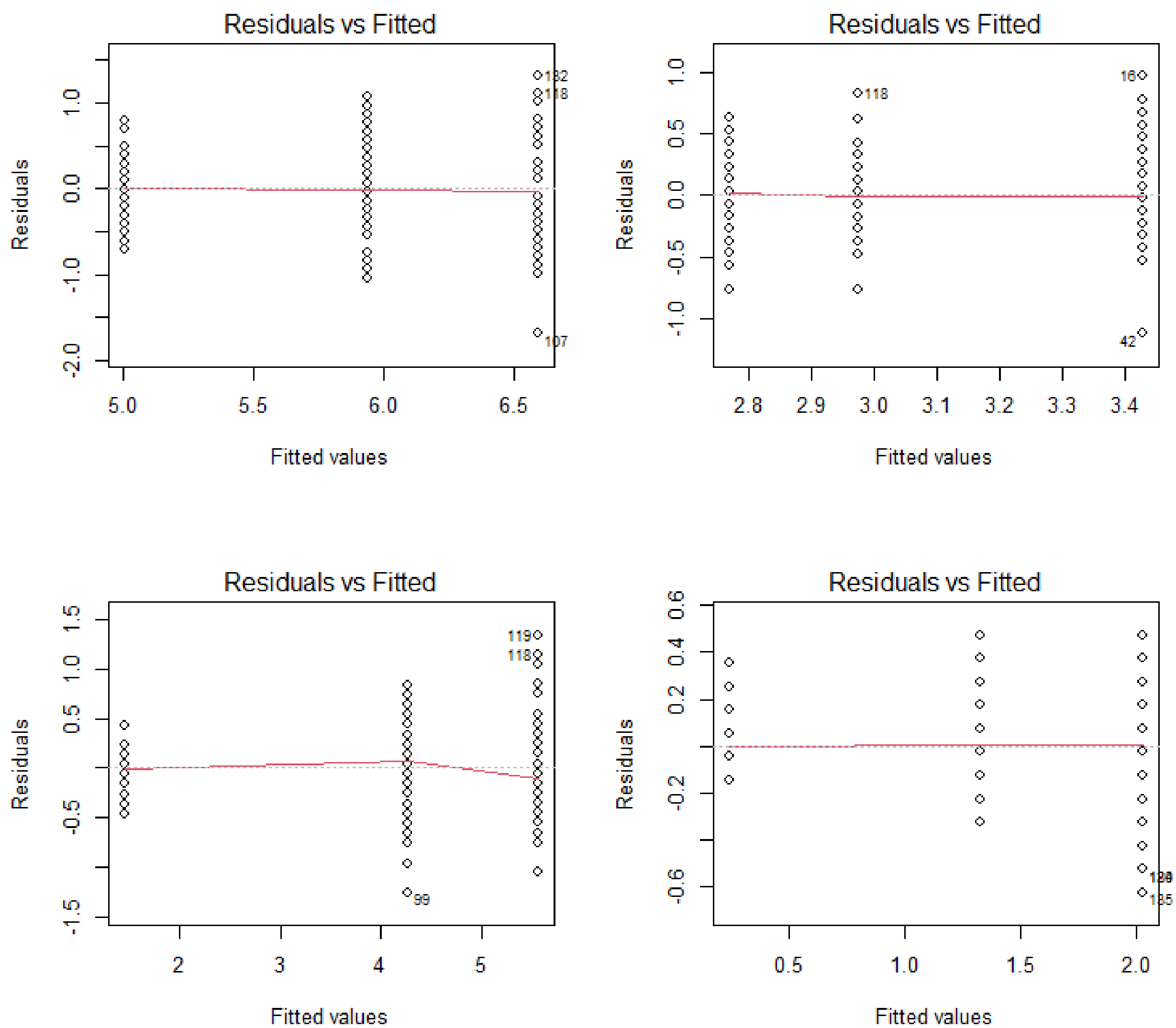
(ii)

Then we perform a one-way ANOVA of the variables Y, X1, X2, X3 on the categorical variable W, the corresponding results are presented bellow:

```
> summary(y_aov)
             Df Sum Sq Mean Sq F value Pr(>F)    
W             2  63.21  31.606   119.3 <2e-16 ***
Residuals   147  38.96   0.265                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(x1_aov)
             Df Sum Sq Mean Sq F value Pr(>F)    
W             2  11.35   5.672   49.16 <2e-16 ***
Residuals   147  16.96   0.115                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(x2_aov)
             Df Sum Sq Mean Sq F value Pr(>F)    
W             2  437.1  218.55    1180 <2e-16 ***
Residuals   147   27.2    0.19                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(x3_aov)
             Df Sum Sq Mean Sq F value Pr(>F)    
W             2  80.41   40.21     960 <2e-16 ***
Residuals   147   6.16    0.04                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(iii)

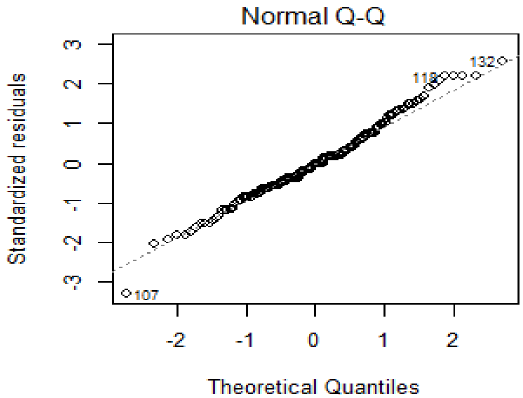
We can realize from the plots and from the p-values at the ANOVA results that the means differ significantly for the different values of the categorical variable W so the variability of all variables Y, X1, X2, X3 is explained by the variable W. At this point its crucial to examine the assumptions of Anova, that is the homogeneity. Also, we have to examine whether our residuals satisfy the Normality. The first depiction bellow checking the homogeneity and with the Bartlett test we calculate the p-values with the null hypothesis of homogeneity and alternative to not be homoscedastic.

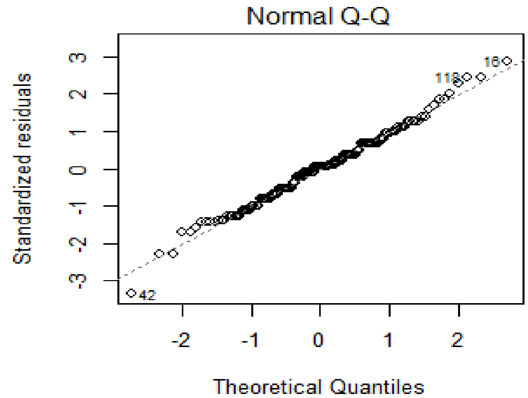


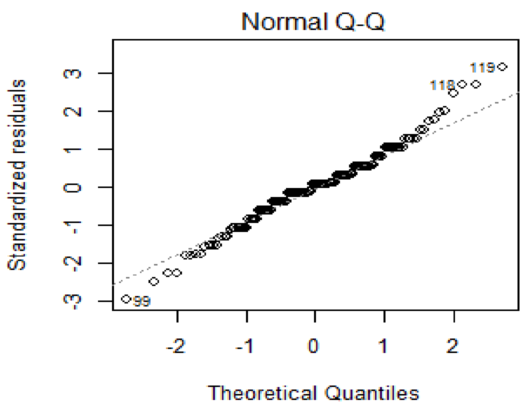
```
> bartlett.test(Y ~ W, data = df)

        Bartlett test of homogeneity of variances

data:  Y by W
Bartlett's K-squared = 16.006, df = 2, p-value = 0.0003345

> bartlett.test(X1 ~ W, data = df)

        Bartlett test of homogeneity of variances

data:  X1 by W
Bartlett's K-squared = 2.0911, df = 2, p-value = 0.3515

> bartlett.test(X2 ~ W, data = df)

        Bartlett test of homogeneity of variances

data:  X2 by W
Bartlett's K-squared = 55.423, df = 2, p-value = 9.229e-13

> bartlett.test(X3 ~ W, data = df)

        Bartlett test of homogeneity of variances

data:  X3 by W
Bartlett's K-squared = 39.213, df = 2, p-value = 3.055e-09
```

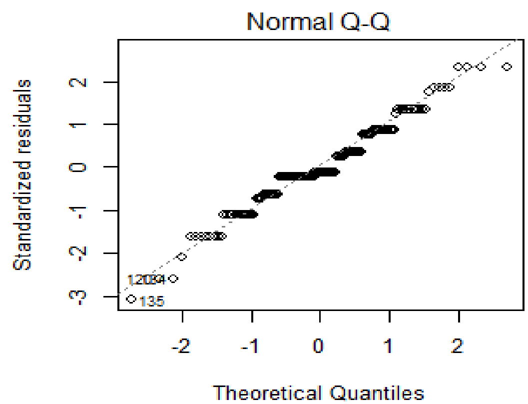
As we can see from the above analysis only the X1 is correct with respect to Anova Assumptions. In all other cases the p-value is less than the a=0.05 level of significance, so we reject our null hypothesis, and hence the homogeneity doesn't hold.

Now for the normality we present the qq plots and we apply Shapiro-Wilk test with null hypothesis of normality and alternative not being normal.





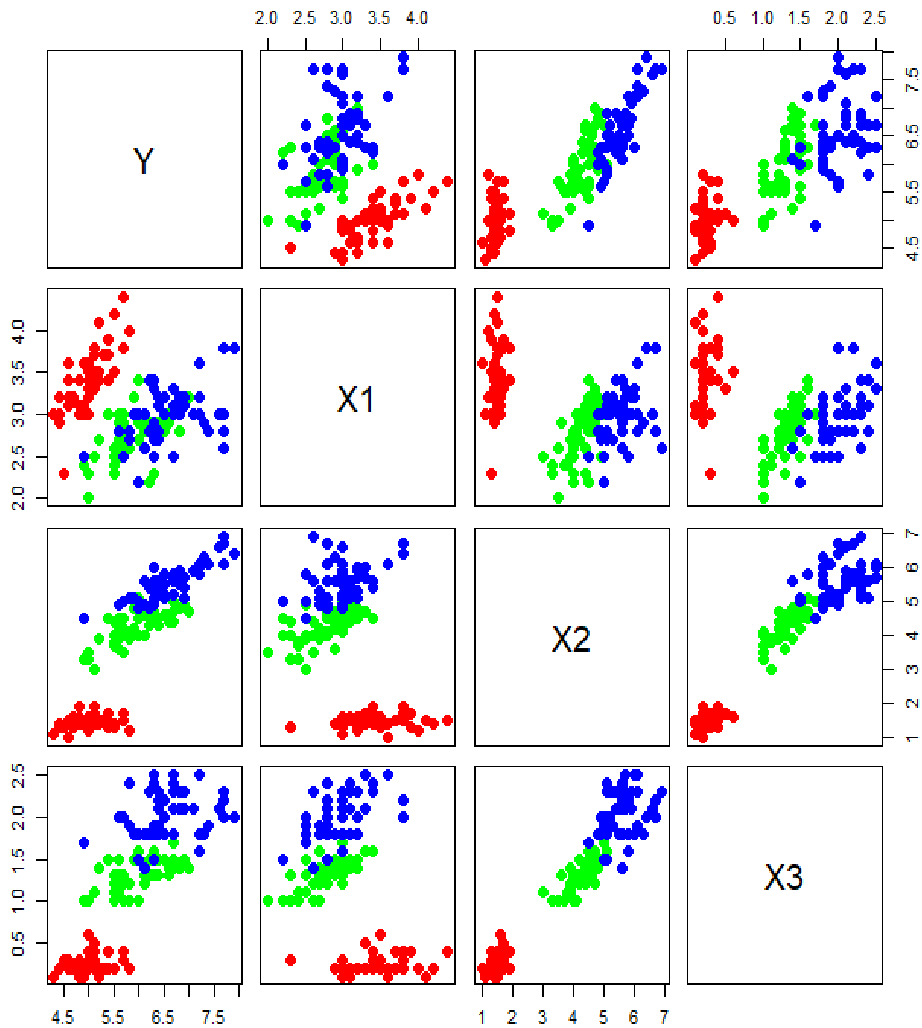




```
> shapiro.test(x = aov_residuals1)

        Shapiro-Wilk normality test

data:  aov_residuals1
W = 0.9879, p-value = 0.2189

> shapiro.test(x = aov_residuals2)

        Shapiro-Wilk normality test

data:  aov_residuals2
W = 0.98948, p-value = 0.323

> shapiro.test(x = aov_residuals3)

        Shapiro-Wilk normality test

data:  aov_residuals3
W = 0.98108, p-value = 0.03676

> shapiro.test(x = aov_residuals4)

        Shapiro-Wilk normality test

data:  aov_residuals4
W = 0.97217, p-value = 0.003866
```

*As we can see, observing the results we get from the tests and the qq plots we can be 95% sure (because our tests conducted with a=0.05 level of significance) that the Anova assumptions are valid for Y and X1. For the rest variables a non-parametric test recommended as Kruskal-Wallis test and also to try to solve the problem of no-homogeneity to take the **Log** of the desired quantities in order to transform them and make them to satisfy the homogeneity assumption.*

(b) *At this part of the exercise we provide a scatterplot matrix of Y,X1,X2,X3.*



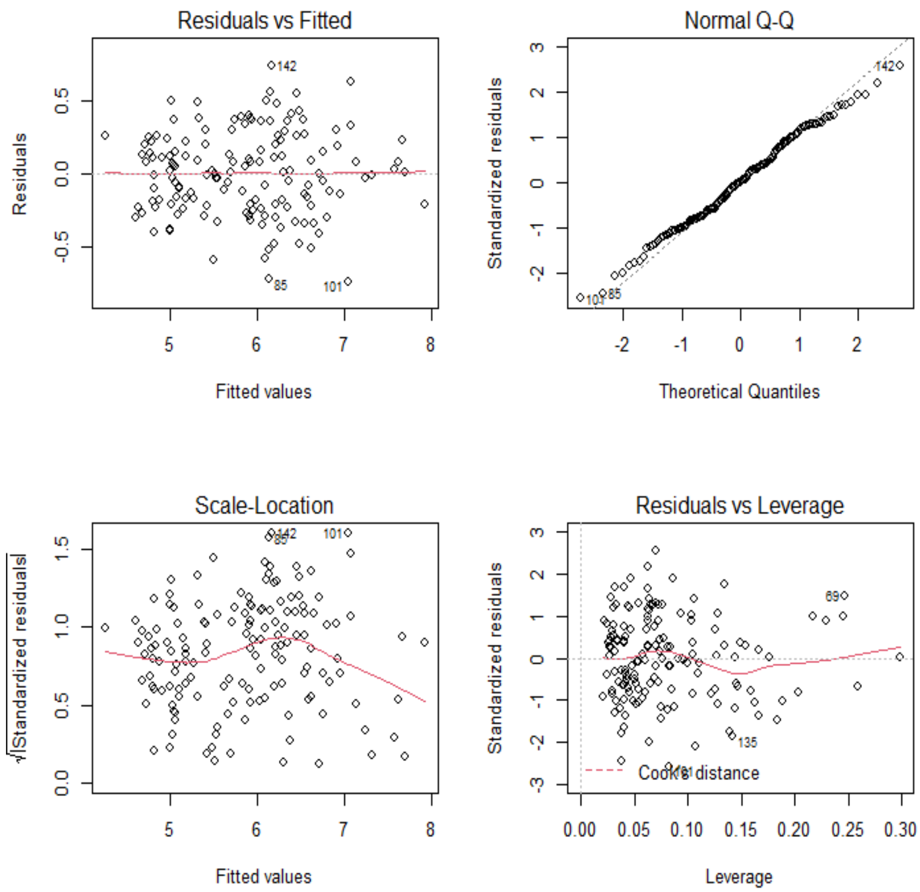
(c) *At this point we run the regression model of Y on X1 coefficients of the model presented bellow:*

```
> model1 = lm(Y~X1, data = df)
> coefficients(model1)
(Intercept)          X1
  6.5262226  -0.2233611
```

(**d)** *Next we run the regression model of Y on all the remaining variables, presented bellow the summary of our model:*

```
lm(formula = Y ~ X1 + X2 + X3 + W + W * X1 + W * X2 + W * X3,
    data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.73883 -0.21607  0.00051  0.21813  0.74427

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.3519     0.4965   4.737 5.33e-06 ***
X1            0.6548     0.1168   5.605 1.09e-07 ***
X2            0.2376     0.2629   0.904   0.3678
X3            0.2521     0.4384   0.575   0.5661
WB           -0.4564     0.6727  -0.678   0.4987
WC           -1.6520     0.7067  -2.338   0.0208 *
X1:WB        -0.2680     0.2172  -1.234   0.2194
X1:WC        -0.3245     0.2016  -1.610   0.1098
X2:WB         0.6708     0.3017   2.223   0.0278 *
X2:WC         0.7080     0.2765   2.561   0.0115 *
X3:WB        -0.9313     0.5866  -1.588   0.1146
X3:WC        -0.4219     0.4765  -0.885   0.3775
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2997 on 138 degrees of freedom
Multiple R-squared:  0.8787,    Adjusted R-squared:  0.869
F-statistic: 90.87 on 11 and 138 DF,  p-value: < 2.2e-16
```

(e) *At this point we examine the regression assumptions.*



From the above depiction its seems that our residuals are close to normal because they spread without pattern across the line(Y). So we can be confident that our residuals doesn't contain useful information.

(f) *At this part of the project we use the "stepwise regression" approach to examine whether we can reduce the dimension of the model.*

```
Start:  AIC=-55.6
Y ~ 1

       Df Sum of Sq     RSS      AIC
+ X2    1    77.643  24.525 -267.641
+ X3    1    68.353  33.815 -219.460
+ W     2    63.212  38.956 -196.230
+ X1    1     1.412 100.756  -55.690
<none>              102.168  -55.602

Step:  AIC=-267.64
Y ~ X2

       Df Sum of Sq     RSS     AIC
+ X1    1     8.196  16.329 -326.66
+ W     2     7.843  16.682 -321.45
+ X3    1     0.644  23.881 -269.63
<none>               24.525 -267.64
- X2    1    77.643 102.168  -55.60

Step:  AIC=-326.66
Y ~ X2 + X1

       Df Sum of Sq     RSS     AIC
+ W     2     2.363  13.966 -346.11
+ X3    1     1.883  14.445 -343.04
<none>               16.329 -326.66
- X1    1     8.196  24.525 -267.64
- X2    1    84.427 100.756  -55.69

Step:  AIC=-346.11
Y ~ X2 + X1 + W

       Df Sum of Sq    RSS     AIC
+ X3    1    0.4090 13.556 -348.57
+ X2:W  2    0.5817 13.384 -348.49
+ X1:W  2    0.5641 13.402 -348.29
<none>              13.966 -346.11
- W     2    2.3632 16.329 -326.66
- X1    1    2.7161 16.682 -321.45
- X2    1   14.0382 28.004 -243.74

Step:  AIC=-348.57
Y ~ X2 + X1 + W + X3

       Df Sum of Sq    RSS     AIC
+ X2:W  2    0.4883 13.068 -350.07
+ X1:W  2    0.3848 13.172 -348.88
+ X3:W  2    0.3720 13.184 -348.74
<none>              13.556 -348.57
- X3    1    0.4090 13.966 -346.11
- W     2    0.8889 14.445 -343.04
- X1    1    3.1250 16.681 -319.45
- X2    1   13.7853 27.342 -245.33

Step:  AIC=-350.07
Y ~ X2 + X1 + W + X3 + X2:W

       Df Sum of Sq    RSS     AIC
+ X1:W  2    0.4398 12.628 -351.20
+ X3:W  2    0.3875 12.681 -350.58
<none>              13.068 -350.07
- X2:W  2    0.4883 13.556 -348.57
- X3    1    0.3156 13.384 -348.49
- X1    1    3.2285 16.297 -318.95

Step:  AIC=-351.2
Y ~ X2 + X1 + W + X3 + X2:W + X1:W

       Df Sum of Sq    RSS     AIC
- X3    1   0.14311 12.771 -351.51
<none>              12.628 -351.20
- X1:W  2   0.43978 13.068 -350.07
+ X3:W  2   0.23370 12.395 -350.00
- X2:W  2   0.54329 13.172 -348.88

Step:  AIC=-351.51
Y ~ X2 + X1 + W + X2:W + X1:W

       Df Sum of Sq    RSS     AIC
<none>              12.771 -351.51
+ X3    1   0.14311 12.628 -351.20
- X1:W  2   0.61227 13.384 -348.49
- X2:W  2   0.62995 13.402 -348.29
```

From the above analysis we conclude that our best model achieve -351.51 AIC, so we qualify the model bellow and we present the coefficients of the model:

```
lm(formula = Y ~ X1 + X2 + W + W * X1 + W * X2, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.79154 -0.20188 -0.00232  0.20058  0.70178 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.3037     0.4915   4.687 6.46e-06 ***
X1            0.6674     0.1153   5.791 4.37e-08 ***
X2            0.2834     0.2516   1.127   0.2618    
WB           -0.1873     0.6581  -0.285   0.7764    
WC           -1.6790     0.6998  -2.399   0.0177 *  
X1:WB        -0.4198     0.2016  -2.082   0.0392 *  
X1:WC        -0.4075     0.1856  -2.195   0.0298 *  
X2:WB         0.4522     0.2748   1.646   0.1021    
X2:WC         0.6514     0.2656   2.453   0.0154 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.301 on 141 degrees of freedom
Multiple R-squared:  0.875,	Adjusted R-squared:  0.8679 
F-statistic: 123.4 on 8 and 141 DF,  p-value: < 2.2e-16
```

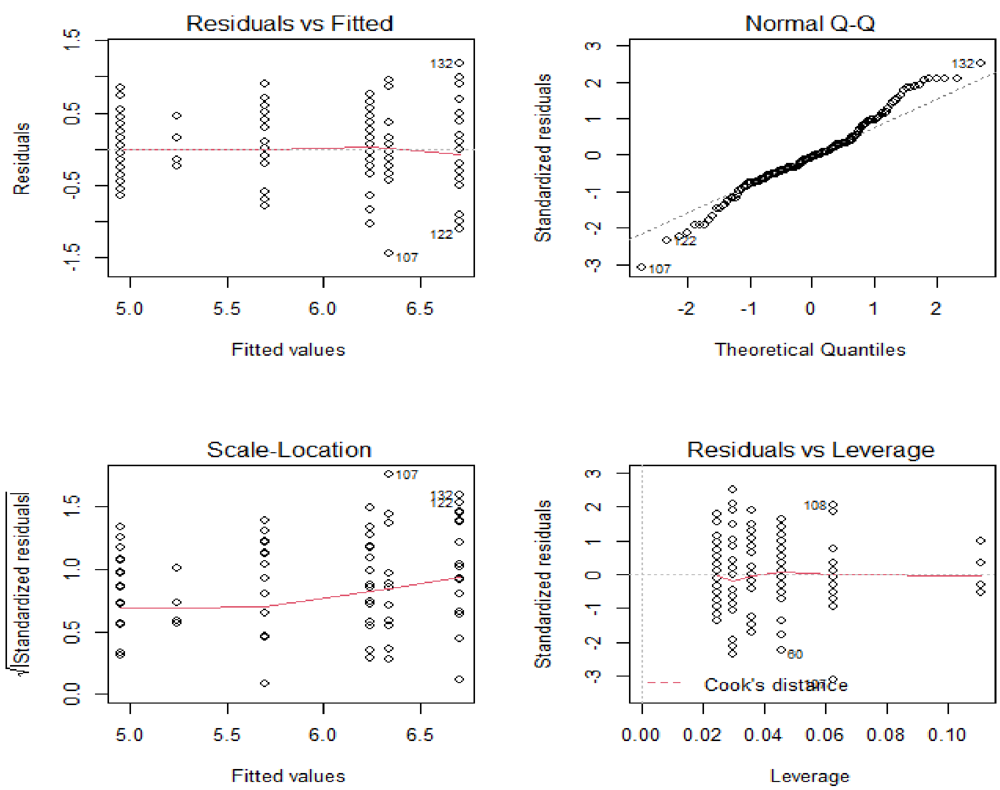
The intercept is the expected value of Y when all predictors are zero and the coefficient bi is the change of the expected value of Y when all predictors are fixed and only the i-th predictor's value increases by 1.

(g) *Afterwards, to construct the Z variable we use quantcut() function in order to create it, with 4 levels based on the quantiles of X3 and we present the contingency table between Z and W.*

```
> table(z,df$W)

z          A  B  C
  25%_X3  41  0  0
  50%_X3   9 28  0
  75%_X3   0 22 16
  100%_X3  0  0 34
```

(h) *Finally, we run the parametric two-way ANOVA of Y on the categorical variables W and Z (including only the main effects). We provide models coefficients and we examine the assumptions.*

```
> anova_two_way <- aov(Y ~ W + z, data = df)
> summary(anova_two_way)
             Df Sum Sq Mean Sq F value   Pr(>F)    
W             2  63.21  31.606 137.073  < 2e-16 ***
z             3   5.75   1.918   8.317 3.87e-05 ***
Residuals   144  33.20   0.231                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> coefficients(anova_two_way)
(Intercept)          WB          WC    z50%_X3    z75%_X3   z100%_X3 
  4.9536585   0.4519841   0.5485750   0.2907859  0.8352664  1.2036488 
```




```
> shapiro.test(two_way_residuals)

        Shapiro-Wilk normality test

data:  two_way_residuals
W = 0.98259, p-value = 0.05473

> lillie.test(two_way_residuals)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  two_way_residuals
D = 0.085335, p-value = 0.009508
```

From the above depiction we realize that the homogeneity seems holding(upper left depiction), regarding normality (observing bottom left and upper right) we conclude that normality seems not holding. Regarding with the p-values above we conclude that W and Z factors above are significant but from examination of residuals we conclude that the analysis is not reliable. Finally observing the Shapiro-Wilk test and Kolmogorov-Smirnov which conducted test about the normality for our residuals we conclude that these 2 test seems to agree about the reliability of our model. The first test indicate p-value =0.05473 meaning that with if we conduct the test with more the 95% level of significance seems that our residuals don't follow normality. The second test (Kolmogorov-Smirnov) shows us that we reject null hypothesis with more certainty meaning that our residuals violate the normality assumptions.