SMT HOMEWORK -6-

Berkay CAYAN

```r
set.seed(123)

X1 = sample(1:100, 100, replace = FALSE)
X2 = sample(101:200, 100, replace = FALSE)


Y = 2 * X1 + 3 * X2 + rnorm(100)


data = data.frame(X1, X2, Y)

head(data)
```

```
  X1  X2        Y
1 31 176 588.7387
2 79 184 710.8375
3 51 146 537.6517
4 14 117 379.6110
5 67 162 619.9521
6 42 198 675.6008
```

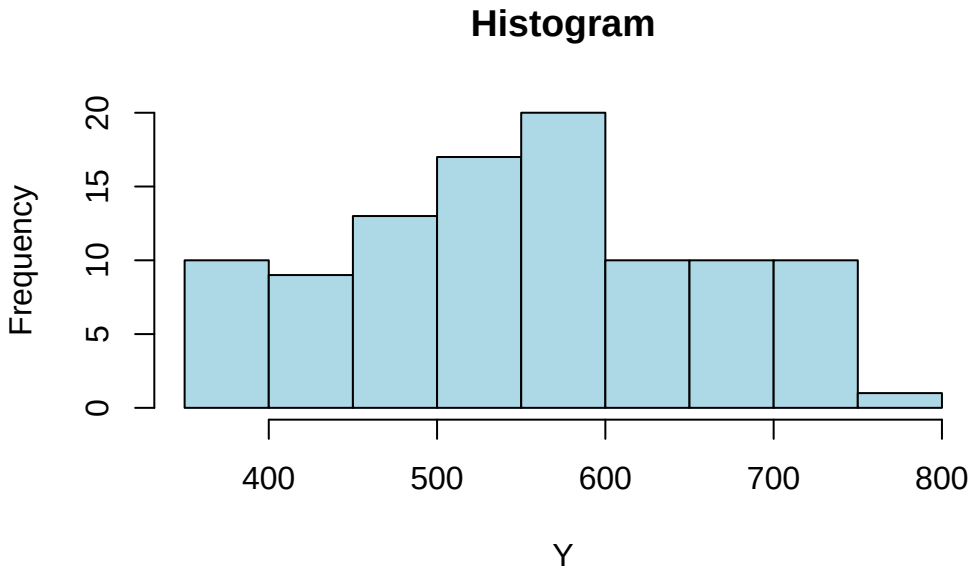
```r
str(data)
```

```
'data.frame':   100 obs. of  3 variables:
 $ X1: int  31 79 51 14 67 42 50 43 97 25 ...
 $ X2: int  176 184 146 117 162 198 154 135 179 124 ...
 $ Y : num  589 711 538 380 620 ...
```

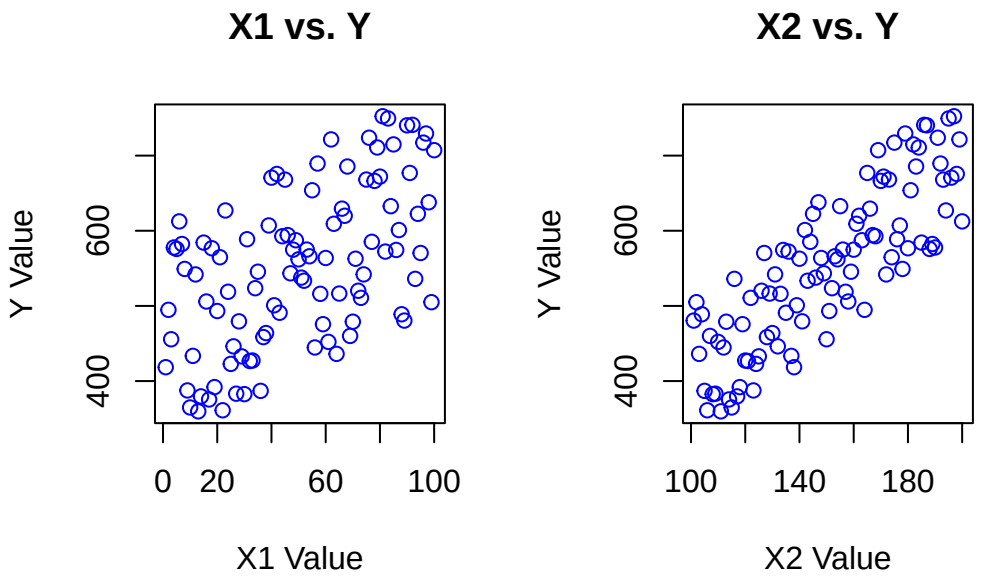
```r
summary(data)
```

```
       X1                X2              Y         
 Min.   :  1.00   Min.   :101.0   Min.   :359.7  
 1st Qu.: 25.75   1st Qu.:125.8   1st Qu.:478.2  
 Median : 50.50   Median :150.5   Median :562.4  
 Mean   : 50.50   Mean   :150.5   Mean   :552.5  
 3rd Qu.: 75.25   3rd Qu.:175.2   3rd Qu.:623.7  
 Max.   :100.00   Max.   :200.0   Max.   :752.4  
```

```r
hist(data$Y, main = "Histogram", xlab = "Y", col = "lightblue")
```



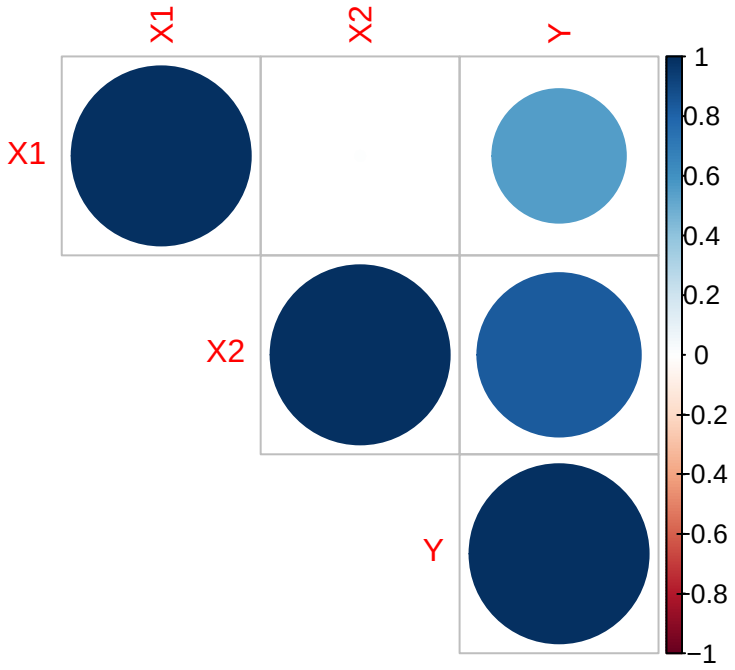

```r
# Scatter plot
par(mfrow = c(1, 2))  # Grafiklerin düzeni
plot(data$X1, data$Y, main = "X1 vs. Y", xlab = "X1 Value", ylab = "Y Value", col = "blue")
plot(data$X2, data$Y, main = "X2 vs. Y", xlab = "X2 Value", ylab = "Y Value", col = "blue")
```




```
cor_matrix <- cor(data)
print(cor_matrix)
```

```
             X1          X2         Y
X1 1.000000000 0.003036304 0.5560641
X2 0.003036304 1.000000000 0.8327551
Y  0.556064080 0.832755121 1.0000000
```

```
corrplot::corrplot(cor_matrix, method = "circle", type = "upper")
```




```r
model <- lm(Y ~ X1 + X2, data = data)

print(summary(model))
```

```
Call:
lm(formula = Y ~ X1 + X2, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4013 -0.7202 -0.0623  0.7411  2.8617

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.750271   0.631299   1.188    0.238
X1          1.995679   0.003916 509.611   <2e-16 ***
X2          2.996271   0.003916 765.119   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.13 on 97 degrees of freedom
Multiple R-squared:  0.9999,    Adjusted R-squared:  0.9999
F-statistic: 4.237e+05 on 2 and 97 DF,  p-value: < 2.2e-16
```

When examining the differences between the values predicted by the model and the actual observations, it is found that the minimum residual is -3.4013 and the maximum residual is 2.8617. The estimated value for the (Intercept) coefficient is 0.750271. However, the p-value for this coefficient is not statistically significant when compared with a significance level of 0.05 ($p > 0.05$). Upon examining the coefficients of the independent variables, the coefficients for X1 and X2 are estimated to be 1.995679 and 2.996271, respectively. Both coefficients are considerably high and statistically significant ($p < 0.05$).

The Residual standard error, which indicates how variable the residuals are, is found to be 1.13. The Multiple R-squared value, which represents the percentage of variance in the dependent variable explained by the independent variables in the model, is quite high at 0.9999, indicating that a large portion of the variance in the dependent variable is explained by the independent variables. The F-statistic, which evaluates the overall statistical significance of the model, is calculated to be 4.237e+05 with a corresponding p-value of < 2.2e-16. These results indicate that the model is overall statistically significant

```r
anova_result <- anova(model)
print(anova_result)
```

```
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 334894  334894  262079 < 2.2e-16 ***
X2         1 748055  748055  585407 < 2.2e-16 ***
Residuals 97    124       1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

X1 and X2 each have one degree of freedom because each is an independent variable. The degrees of freedom for residuals are found by subtracting the number of independent variables (100) from the total number of observations, which is 97 in this case. This table shows the sum of squares for each effect (X1, X2) and for residuals, indicating the total variance. The mean square for each effect is obtained by dividing the variance of each effect by its degrees of freedom. For example, the Mean Sq for X1 is calculated as Sum Sq / Df (334894 / 1). This is the variance analysis F statistic for X1 and X2. The F value indicates how much of the variance in the model is explained. High F values suggest a good fit of the model to the data, while low F values may indicate a poor fit. This indicates the statistical significance of the F value. If it is typically less than 0.05, the respective effect is considered significant in the model. Here, the p values for both effects are very small (Pr(>F) < 2.2e-16), indicating that X1 and X2 are significant in the model. These are the symbols used to indicate the statistical significance of p values. For example, ' ' means $p < 0.05$, ' ' means $p < 0.01$, and ' ' means $p <$

0.001. According to this output, it can be said that X1 and X2 are important for explaining Y and that the model generally fits the data.

```r
ancova_result <- aov(Y ~ X1 + X2, data = data)
print(summary(ancova_result))
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
X1           1 334894  334894  262079 <2e-16 ***
X2           1 748055  748055  585407 <2e-16 ***
Residuals   97    124       1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Represents the degrees of freedom associated with each predictor variable (X1 and X2) and the residuals. Indicates the sum of squares for each predictor variable and the residuals. The mean square is obtained by dividing the sum of squares by its respective degrees of freedom, reflecting the variance explained by each predictor variable and the residuals. The F value is a test statistic for ANCOVA, measuring the ratio of the variance explained by the model to the residual variance, assessing whether there are significant differences among the group means. Represents the p-value associated with each predictor variable, indicating the statistical significance of each variable in the model. In this output, both X1 and X2 have extremely small p-values (less than 2e-16), providing strong evidence against the null hypothesis and suggesting that they are highly significant predictors of the response variable Y. These codes offer a quick reference to the significance levels of the predictors. In this case, '***' denotes $p < 0.001$, indicating extreme significance.

In summary, both X1 and X2 are highly significant predictors of the response variable Y in the ANCOVA model. The model appears to fit the data well, as evidenced by the highly significant F values.