A2 - Regresión Múltiple

Héctor Hibran Tapia Fernández - A01661114

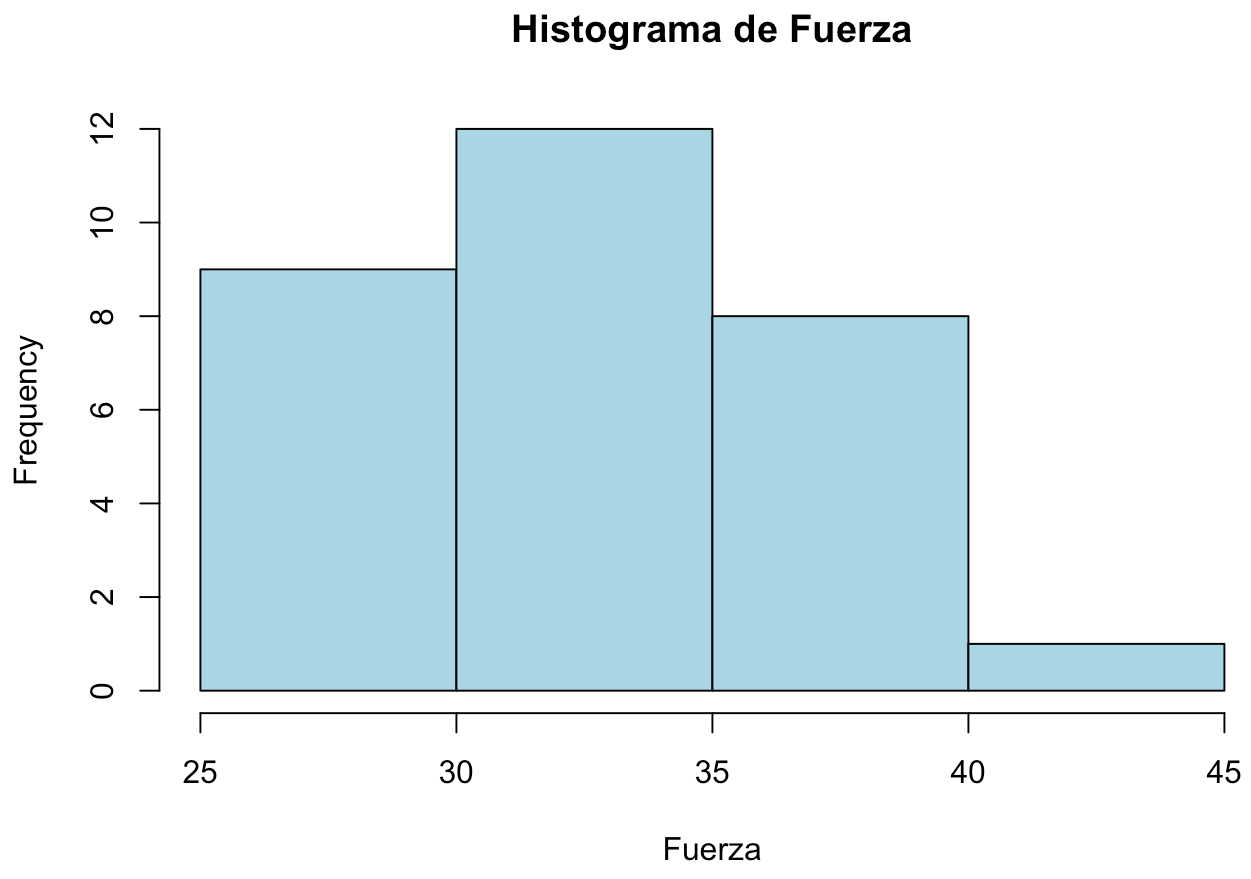
2024-09-20

1. Haz un análisis descriptivo de los datos: medidas principales y gráficos

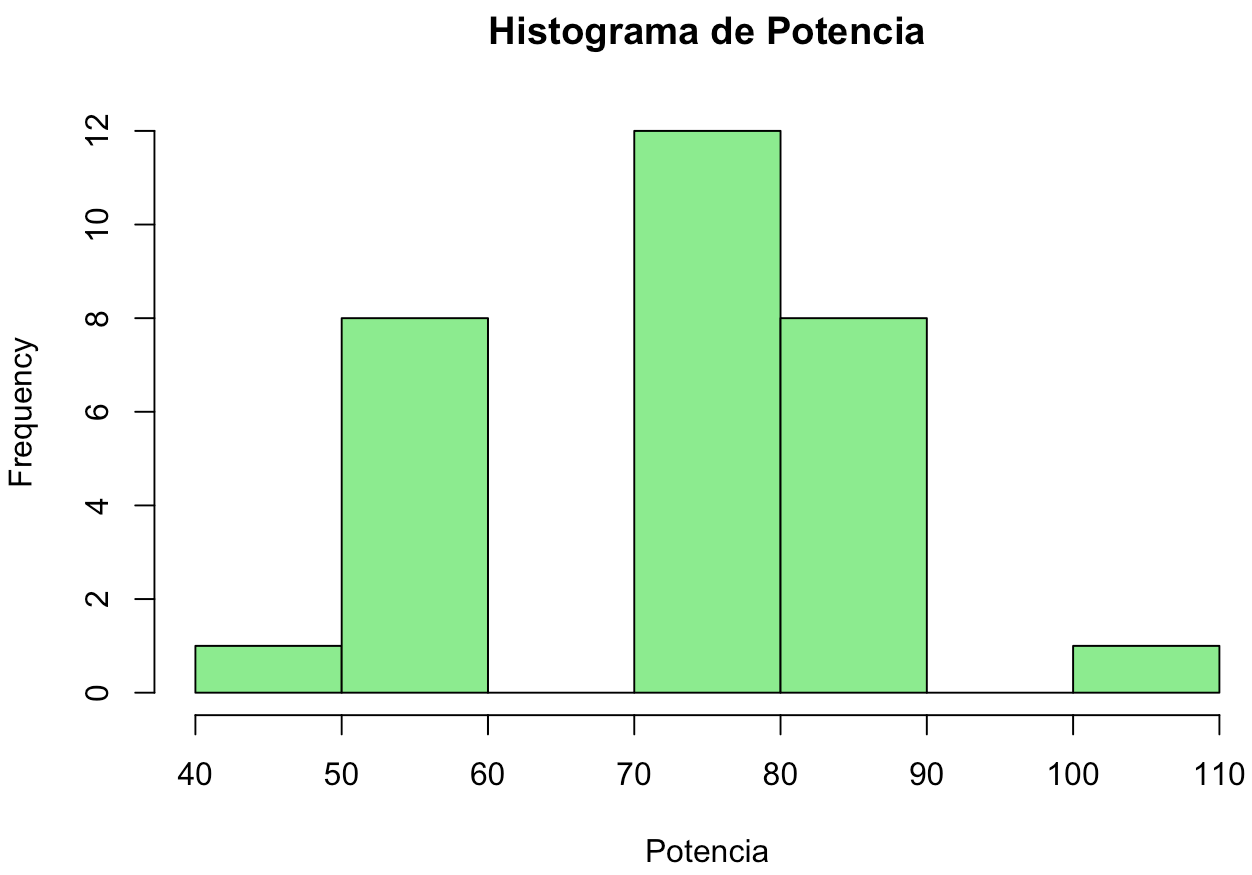
```
df = read.csv("./AlCorte.csv")
summary(df)
```

```
##      Fuerza      Potencia    Temperatura      Tiempo    Resistencia
##  Min.   :25   Min.   : 45   Min.   :150   Min.   :10   Min.   :22.70
##  1st Qu.:30   1st Qu.: 60   1st Qu.:175   1st Qu.:15   1st Qu.:34.67
##  Median :35   Median : 75   Median :200   Median :20   Median :38.60
##  Mean   :35   Mean   : 75   Mean   :200   Mean   :20   Mean   :38.41
##  3rd Qu.:40   3rd Qu.: 90   3rd Qu.:225   3rd Qu.:25   3rd Qu.:42.70
##  Max.   :45   Max.   :105   Max.   :250   Max.   :30   Max.   :58.70
```

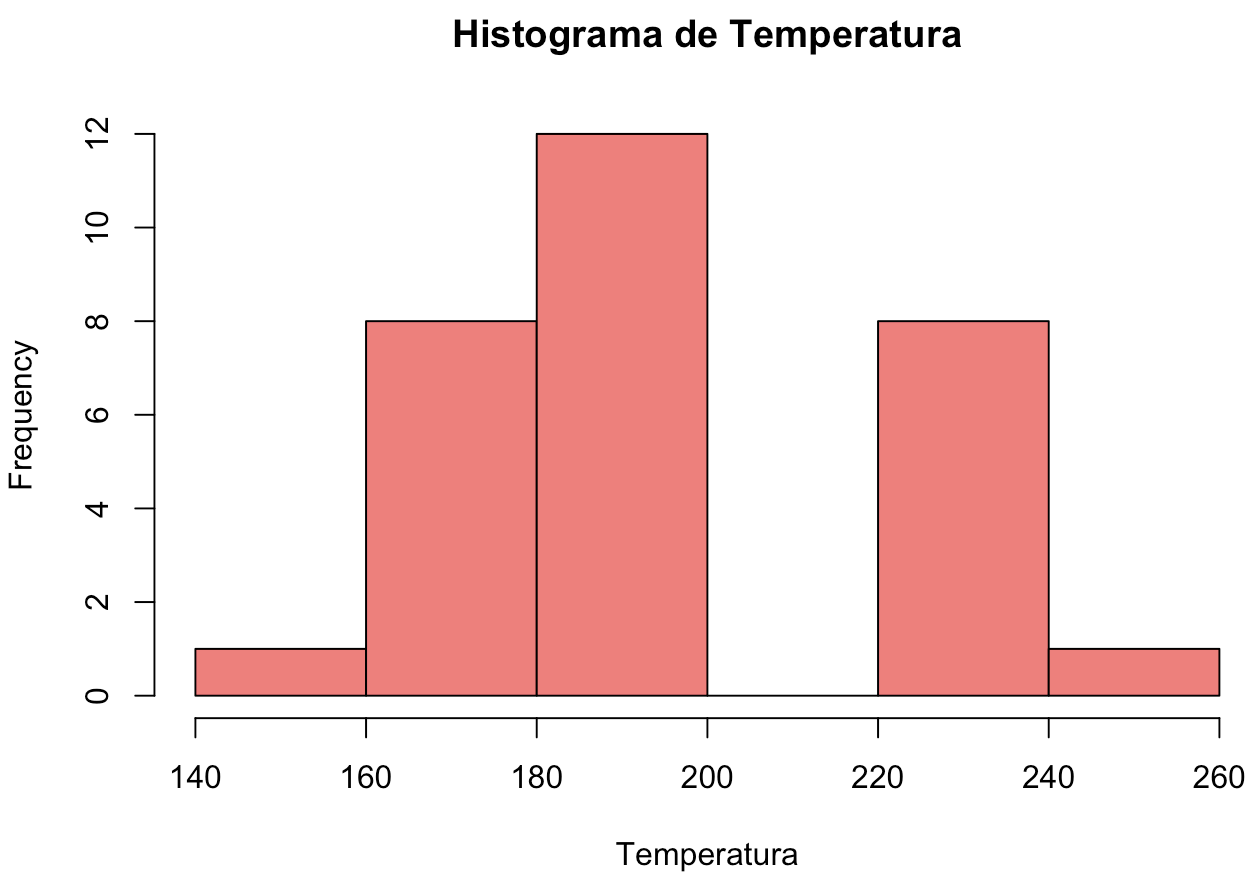
```
hist(df$Fuerza, main = "Histograma de Fuerza", xlab = "Fuerza", col = "lightblue")
```



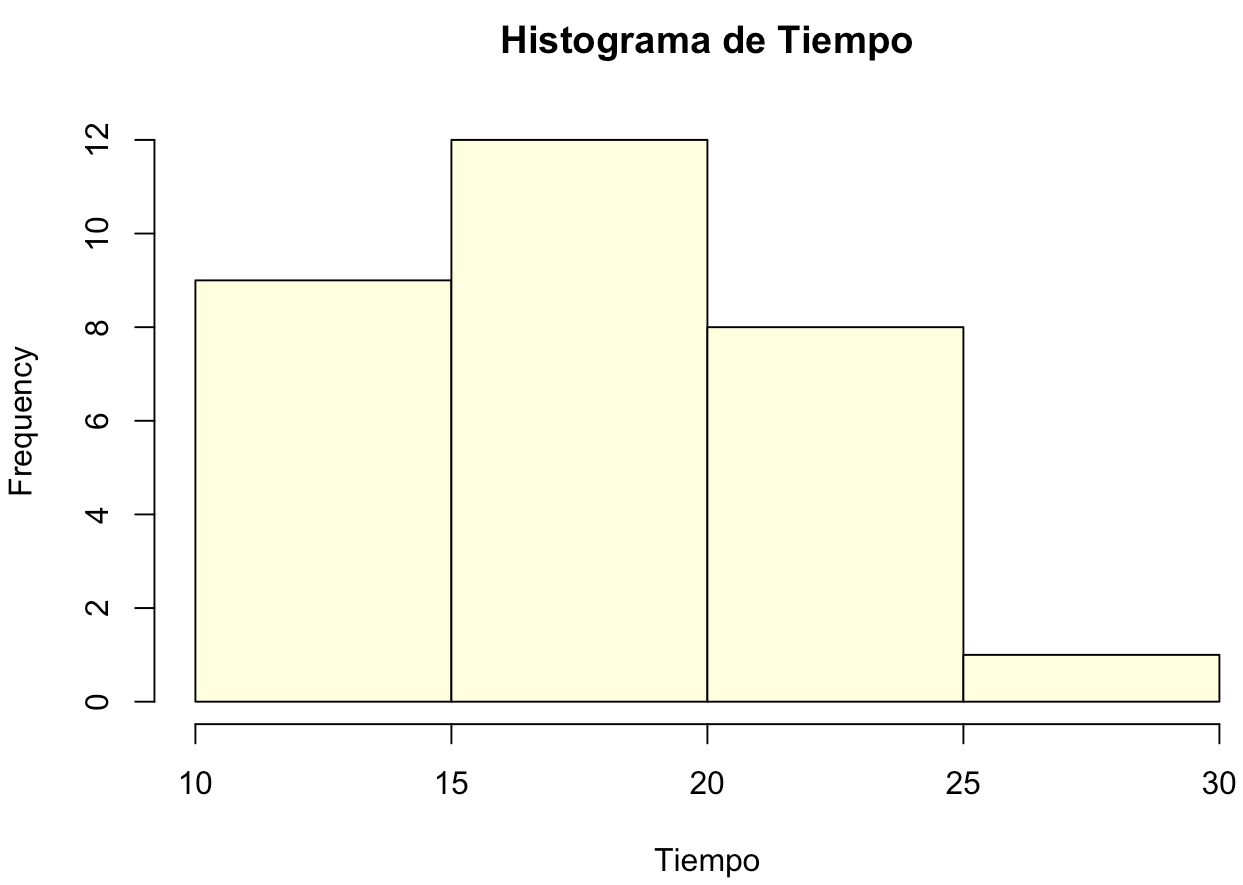
```
hist(df$Potencia, main = "Histograma de Potencia", xlab = "Potencia", col = "lightgree
n")
```



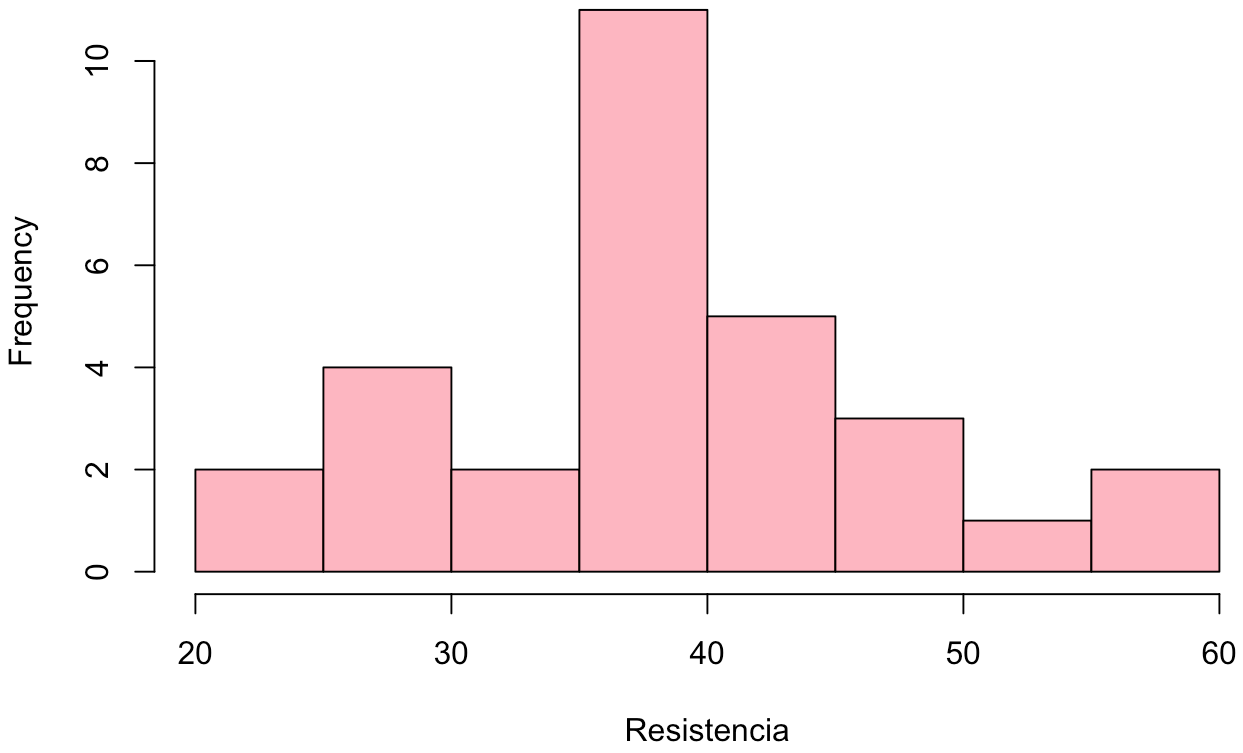

```
hist(df$Temperatura, main = "Histograma de Temperatura", xlab = "Temperatura", col = "li
ghtcoral")
```



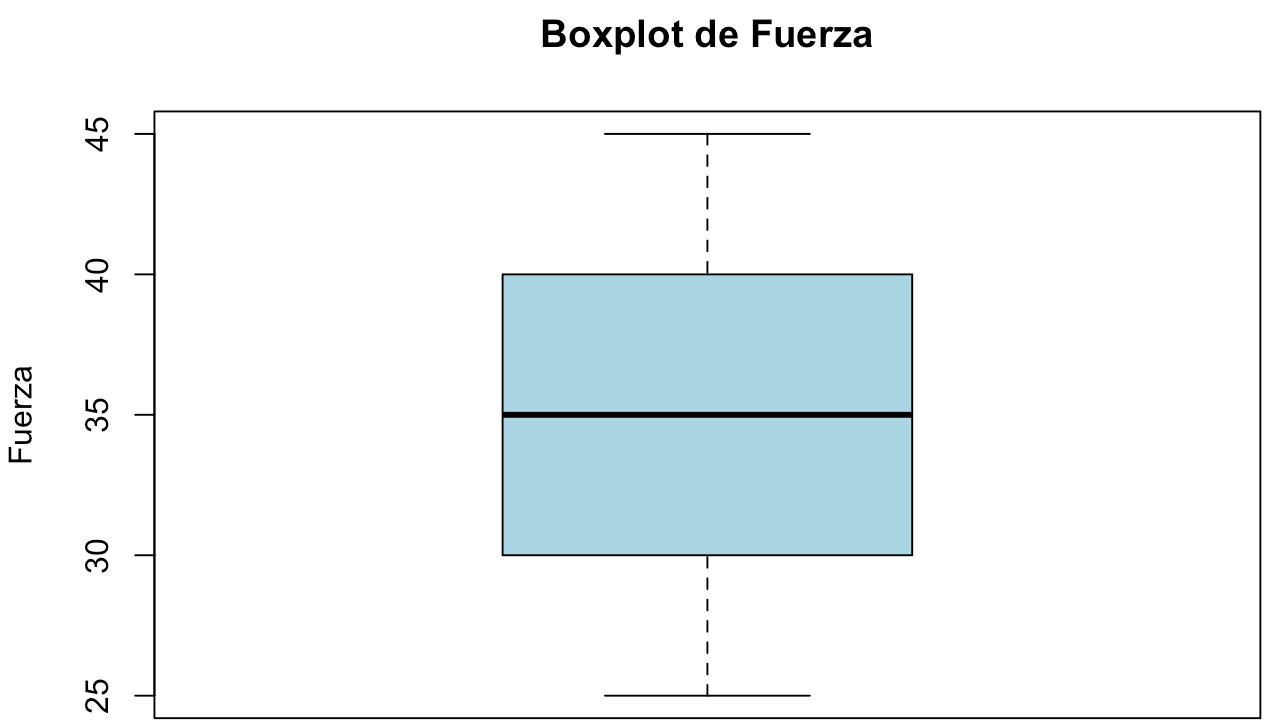

```
hist(df$Tiempo, main = "Histograma de Tiempo", xlab = "Tiempo", col = "lightyellow")
```



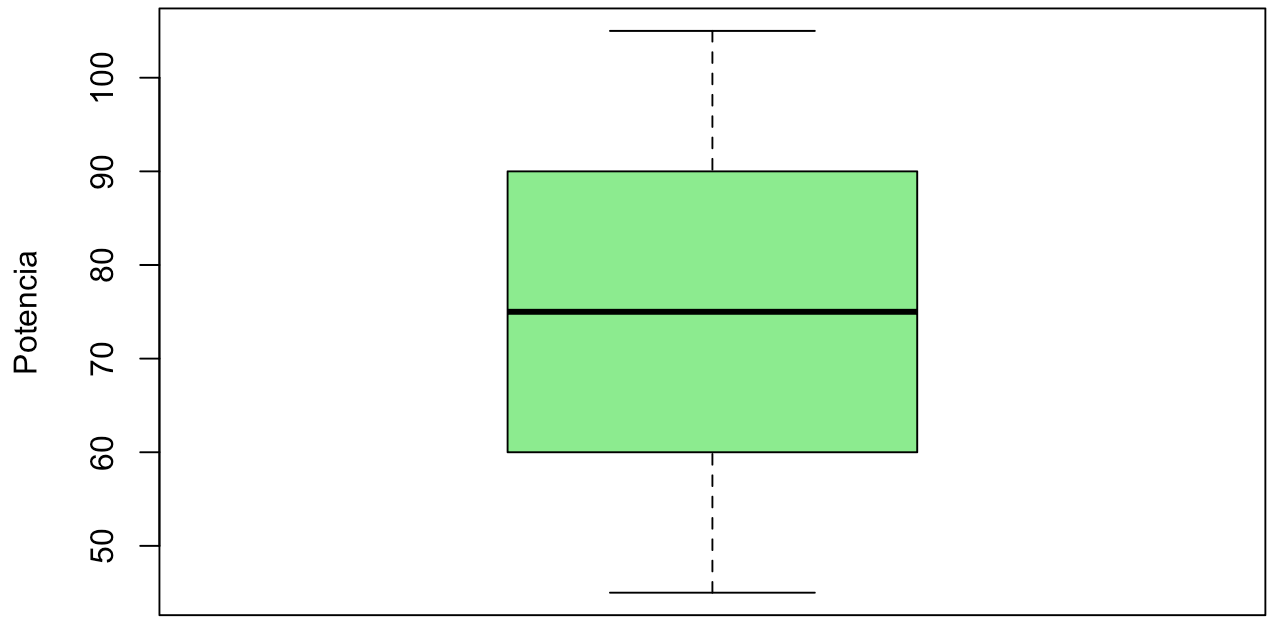

```
hist(df$Resistencia, main = "Histograma de Resistencia", xlab = "Resistencia", col = "li
ghtpink")
```



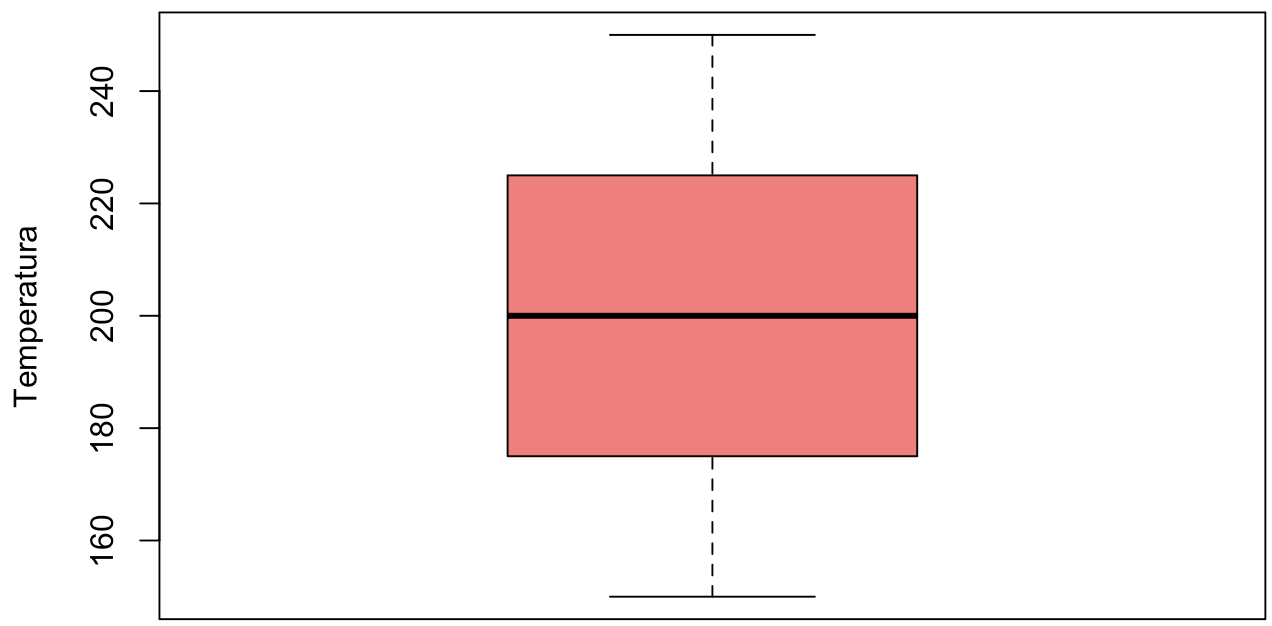

```
boxplot(df$Fuerza, main = "Boxplot de Fuerza", col = "lightblue", ylab = "Fuerza")
```



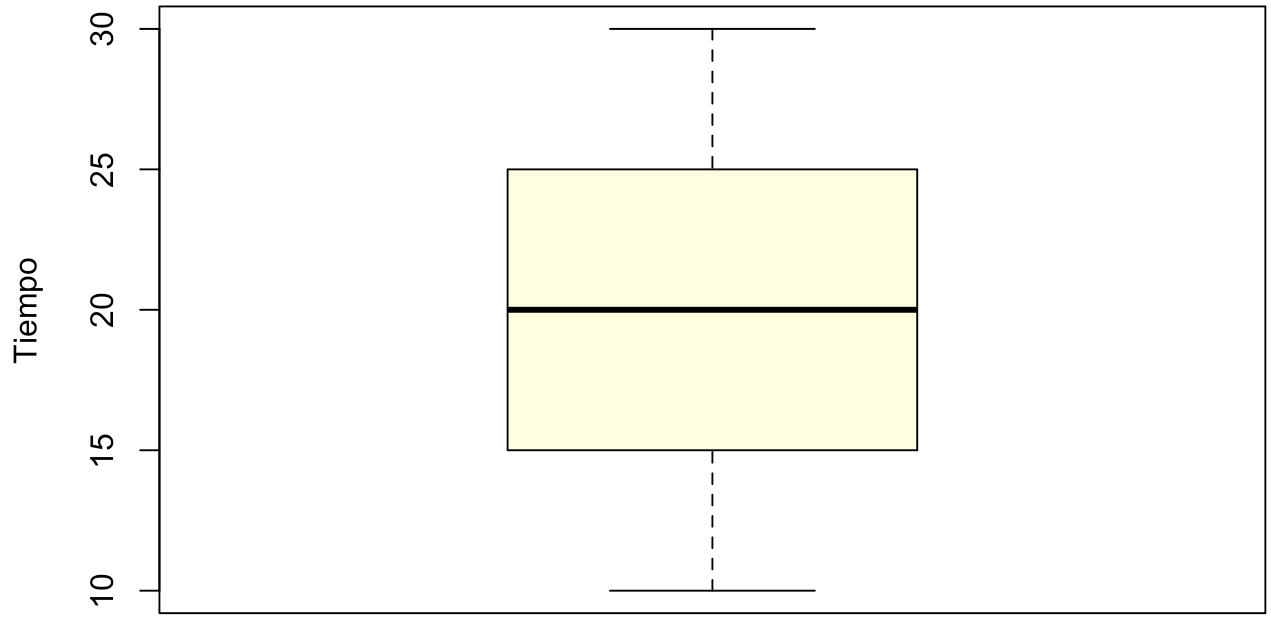

```
boxplot(df$Potencia, main = "Boxplot de Potencia", col = "lightgreen", ylab = "Potenci
a")
```



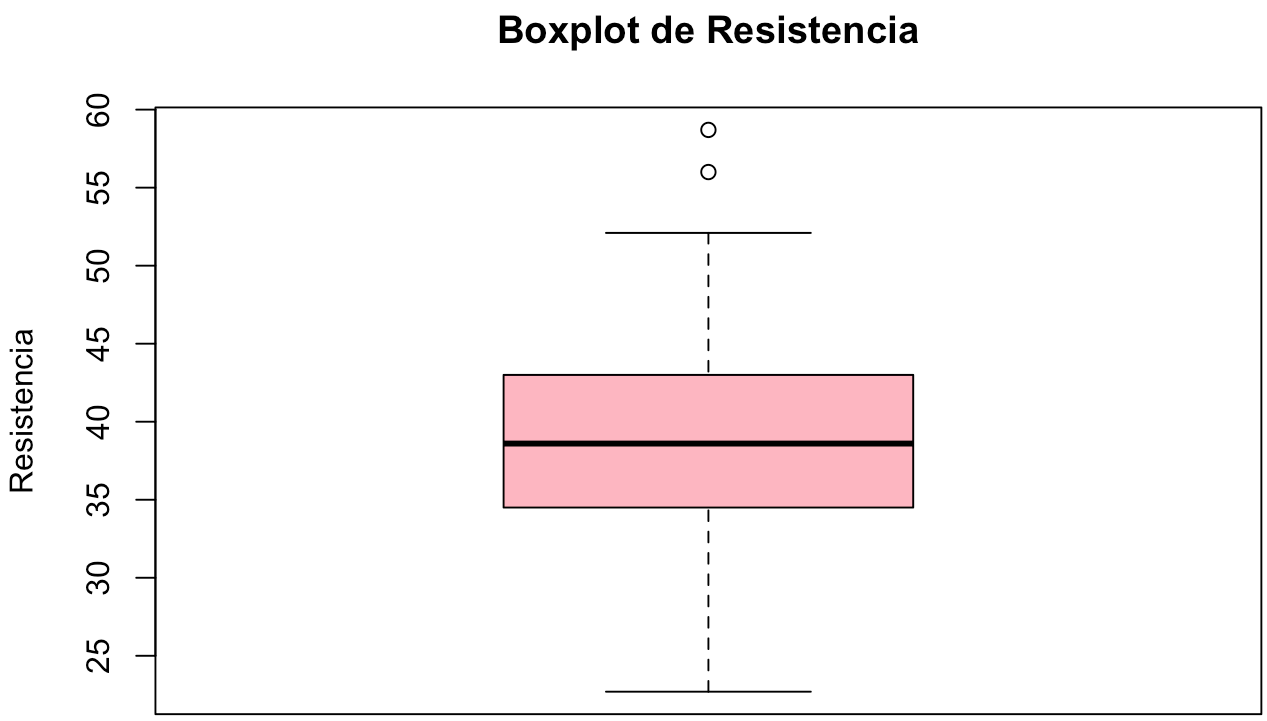

```
boxplot(df$Temperatura, main = "Boxplot de Temperatura", col = "lightcoral", ylab = "Tem
peratura")
```



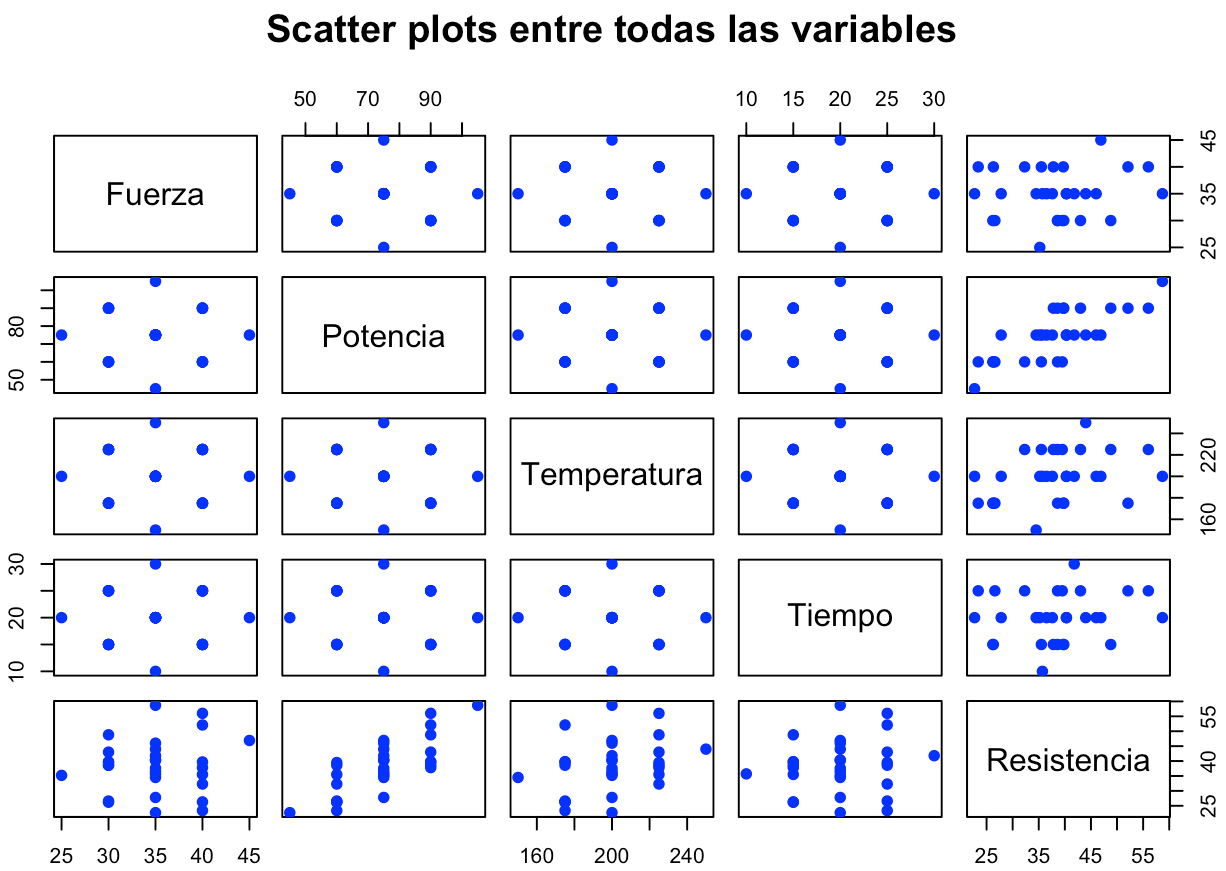

```
boxplot(df$Tiempo, main = "Boxplot de Tiempo", col = "lightyellow", ylab = "Tiempo")
```




```
boxplot(df$Resistencia, main = "Boxplot de Resistencia", col = "lightpink", ylab = "Resi
stencia")
```




```
pairs(df, main = "Scatter plots entre todas las variables", col = "blue", pch = 19)
```



2. Encuentra el mejor modelo de regresión que explique la variable Resistencia.

```
modelo_completo = lm(Resistencia ~ ., data = df)
modelo_nulo = lm(Resistencia ~ 1, data = df)

# Modelo Mixto
modelo_mixto = step(modelo_completo, direction = "both", trace = 1)
```

```
## Start:  AIC=102.96
## Resistencia ~ Fuerza + Potencia + Temperatura + Tiempo
##
##               Df Sum of Sq     RSS    AIC
## - Fuerza       1     26.88  692.00 102.15
## - Tiempo       1     40.04  705.16 102.72
## <none>                      665.12 102.96
## - Temperatura  1    252.20  917.32 110.61
## - Potencia     1   1341.01 2006.13 134.08
##
## Step:  AIC=102.15
## Resistencia ~ Potencia + Temperatura + Tiempo
##
##               Df Sum of Sq     RSS    AIC
## - Tiempo       1     40.04  732.04 101.84
## <none>                      692.00 102.15
## + Fuerza       1     26.88  665.12 102.96
## - Temperatura  1    252.20  944.20 109.47
## - Potencia     1   1341.01 2033.02 132.48
##
## Step:  AIC=101.84
## Resistencia ~ Potencia + Temperatura
##
##               Df Sum of Sq     RSS    AIC
## <none>                      732.04 101.84
## + Tiempo       1     40.04  692.00 102.15
## + Fuerza       1     26.88  705.16 102.72
## - Temperatura  1    252.20  984.24 108.72
## - Potencia     1   1341.01 2073.06 131.07
```

```
# Modelo Forward
modelo_forward = step(modelo_nulo, scope = list(lower = modelo_nulo, upper = modelo_comp
leto), direction = "forward")
```

```
## Start:  AIC=132.51
## Resistencia ~ 1
##
##               Df Sum of Sq     RSS    AIC
## + Potencia     1   1341.01  984.24 108.72
## + Temperatura  1    252.20 2073.06 131.07
## <none>                     2325.26 132.51
## + Tiempo       1     40.04 2285.22 133.99
## + Fuerza       1     26.88 2298.38 134.16
##
## Step:  AIC=108.72
## Resistencia ~ Potencia
##
##               Df Sum of Sq    RSS    AIC
## + Temperatura  1   252.202 732.04 101.84
## <none>                     984.24 108.72
## + Tiempo       1    40.042 944.20 109.47
## + Fuerza       1    26.882 957.36 109.89
##
## Step:  AIC=101.84
## Resistencia ~ Potencia + Temperatura
##
##          Df Sum of Sq    RSS    AIC
## <none>                 732.04 101.84
## + Tiempo  1    40.042 692.00 102.15
## + Fuerza  1    26.882 705.16 102.72
```

```
# Modelo Backward
modelo_backward = step(modelo_completo, direction = "backward")
```

```
## Start:  AIC=102.96
## Resistencia ~ Fuerza + Potencia + Temperatura + Tiempo
##
##               Df Sum of Sq     RSS    AIC
## - Fuerza       1     26.88  692.00 102.15
## - Tiempo       1     40.04  705.16 102.72
## <none>                      665.12 102.96
## - Temperatura  1    252.20  917.32 110.61
## - Potencia     1   1341.01 2006.13 134.08
##
## Step:  AIC=102.15
## Resistencia ~ Potencia + Temperatura + Tiempo
##
##               Df Sum of Sq     RSS    AIC
## - Tiempo       1     40.04  732.04 101.84
## <none>                      692.00 102.15
## - Temperatura  1    252.20  944.20 109.47
## - Potencia     1   1341.01 2033.02 132.48
##
## Step:  AIC=101.84
## Resistencia ~ Potencia + Temperatura
##
##               Df Sum of Sq     RSS    AIC
## <none>                      732.04 101.84
## - Temperatura  1     252.2  984.24 108.72
## - Potencia     1    1341.0 2073.06 131.07
```

Analiza el modelo basándote en:

- Significancia del modelo:

1. Economía de las variables

```
summary(modelo_mixto)
```

```
##
## Call:
## lm(formula = Resistencia ~ Potencia + Temperatura, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.3233  -2.8067  -0.8483   3.1892   9.4600
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -24.90167   10.07207  -2.472  0.02001 *
## Potencia      0.49833    0.07086   7.033 1.47e-07 ***
## Temperatura   0.12967    0.04251   3.050  0.00508 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.207 on 27 degrees of freedom
## Multiple R-squared:  0.6852, Adjusted R-squared:  0.6619
## F-statistic: 29.38 on 2 and 27 DF,  p-value: 1.674e-07
```

```
cat("-----------------------------------------------------------------")
```

```
## -----------------------------------------------------------------
```

```
summary(modelo_forward)
```

```
##
## Call:
## lm(formula = Resistencia ~ Potencia + Temperatura, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.3233  -2.8067  -0.8483   3.1892   9.4600
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -24.90167   10.07207  -2.472  0.02001 *
## Potencia      0.49833    0.07086   7.033 1.47e-07 ***
## Temperatura   0.12967    0.04251   3.050  0.00508 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.207 on 27 degrees of freedom
## Multiple R-squared:  0.6852, Adjusted R-squared:  0.6619
## F-statistic: 29.38 on 2 and 27 DF,  p-value: 1.674e-07
```

```
cat("-----------------------------------------------------------------")
```

```
## ------------------------------------------------------------------
```

```
summary(modelo_backward)
```

```
##
## Call:
## lm(formula = Resistencia ~ Potencia + Temperatura, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.3233  -2.8067  -0.8483   3.1892   9.4600
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -24.90167   10.07207  -2.472  0.02001 *
## Potencia      0.49833    0.07086   7.033 1.47e-07 ***
## Temperatura   0.12967    0.04251   3.050  0.00508 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.207 on 27 degrees of freedom
## Multiple R-squared:  0.6852, Adjusted R-squared:  0.6619
## F-statistic: 29.38 on 2 and 27 DF,  p-value: 1.674e-07
```

2. Significación global (Prueba para el modelo)

```
summary(modelo_mixto)$fstatistic
```

```
##    value    numdf    dendf
## 29.38141  2.00000 27.00000
```

```
print("--------------------------")
```

```
## [1] "--------------------------"
```

```
summary(modelo_forward)$fstatistic
```

```
##    value    numdf    dendf
## 29.38141  2.00000 27.00000
```

```
print("--------------------------")
```

```
## [1] "--------------------------"
```

```
summary(modelo_backward)$fstatistic
```

```
##    value    numdf    dendf
## 29.38141  2.00000 27.00000
```

3. Significación individual (Prueba para cada βi)

```
summary(modelo_mixto)$coefficients
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) -24.9016667 10.07206836 -2.472349 2.001412e-02
## Potencia      0.4983333  0.07085806  7.032839 1.465430e-07
## Temperatura   0.1296667  0.04251483  3.049916 5.082118e-03
```

```
confint(modelo_mixto)
```

```
##                    2.5 %     97.5 %
## (Intercept) -45.56784390 -4.2354894
## Potencia      0.35294461  0.6437221
## Temperatura   0.04243343  0.2168999
```

```
print("------------------------------------------------------------")
```

```
## [1] "------------------------------------------------------------"
```

```
summary(modelo_forward)$coefficients
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) -24.9016667 10.07206836 -2.472349 2.001412e-02
## Potencia      0.4983333  0.07085806  7.032839 1.465430e-07
## Temperatura   0.1296667  0.04251483  3.049916 5.082118e-03
```

```
confint(modelo_forward)
```

```
##                    2.5 %     97.5 %
## (Intercept) -45.56784390 -4.2354894
## Potencia      0.35294461  0.6437221
## Temperatura   0.04243343  0.2168999
```

```
print("------------------------------------------------------------")
```

```
## [1] "------------------------------------------------------------"
```

```
summary(modelo_backward)$coefficients
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) -24.9016667 10.07206836 -2.472349 2.001412e-02
## Potencia      0.4983333  0.07085806  7.032839 1.465430e-07
## Temperatura   0.1296667  0.04251483  3.049916 5.082118e-03
```

```
confint(modelo_backward)
```

```
##                    2.5 %     97.5 %
## (Intercept) -45.56784390 -4.2354894
## Potencia      0.35294461  0.6437221
## Temperatura   0.04243343  0.2168999
```

4. Variación explicada por el modelo

```
summary(modelo_mixto)$r.squared
```

```
## [1] 0.6851783
```

```
summary(modelo_mixto)$adj.r.squared
```

```
## [1] 0.6618581
```

```
print("----------")
```

```
## [1] "----------"
```

```
summary(modelo_forward)$r.squared
```

```
## [1] 0.6851783
```

```
summary(modelo_forward)$adj.r.squared
```

```
## [1] 0.6618581
```

```
print("----------")
```

```
## [1] "----------"
```

```
summary(modelo_backward)$r.squared
```

```
## [1] 0.6851783
```

```
summary(modelo_backward)$adj.r.squared
```

```
## [1] 0.6618581
```

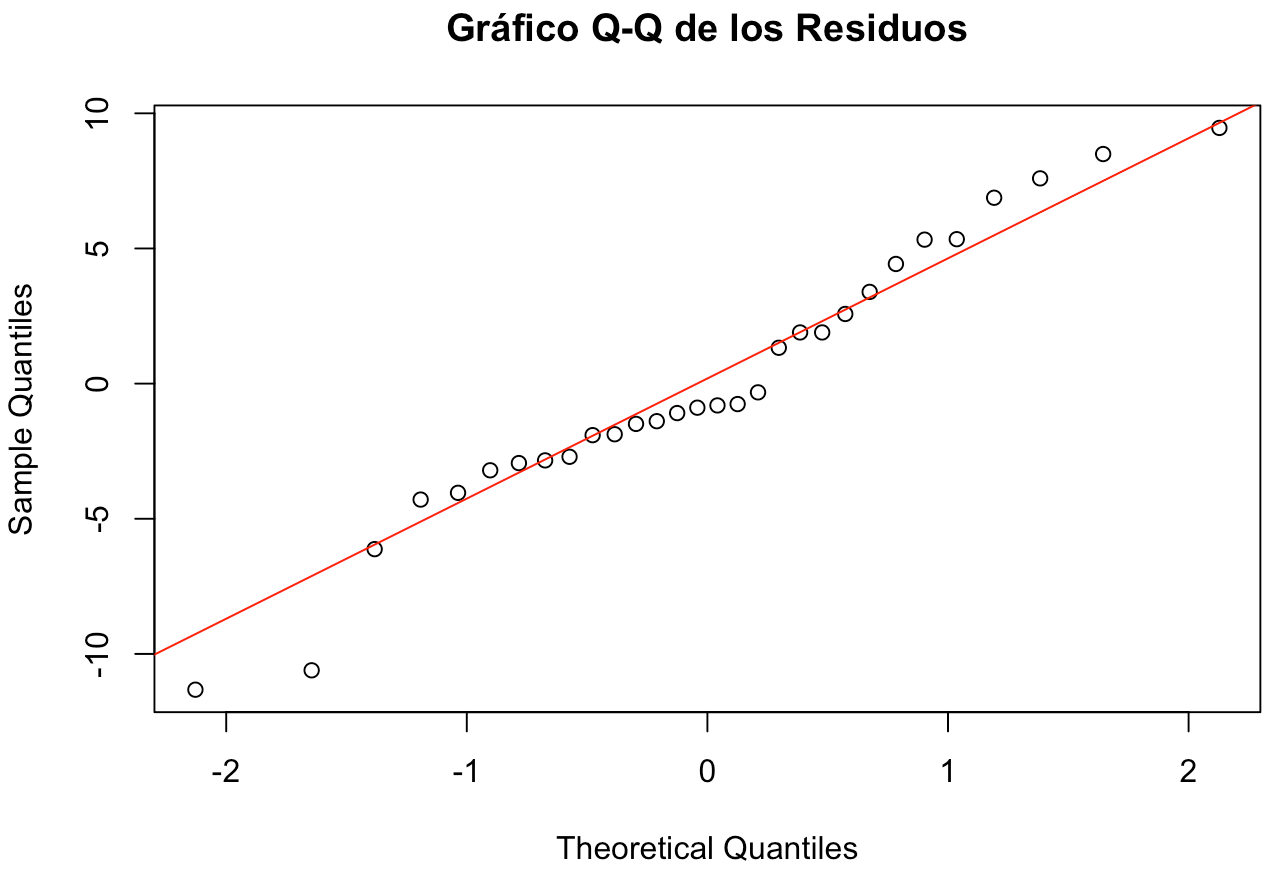
3. Analiza la validez del modelo encontrado:

- Análisis de residuos (homocedasticidad, independencia, etc)

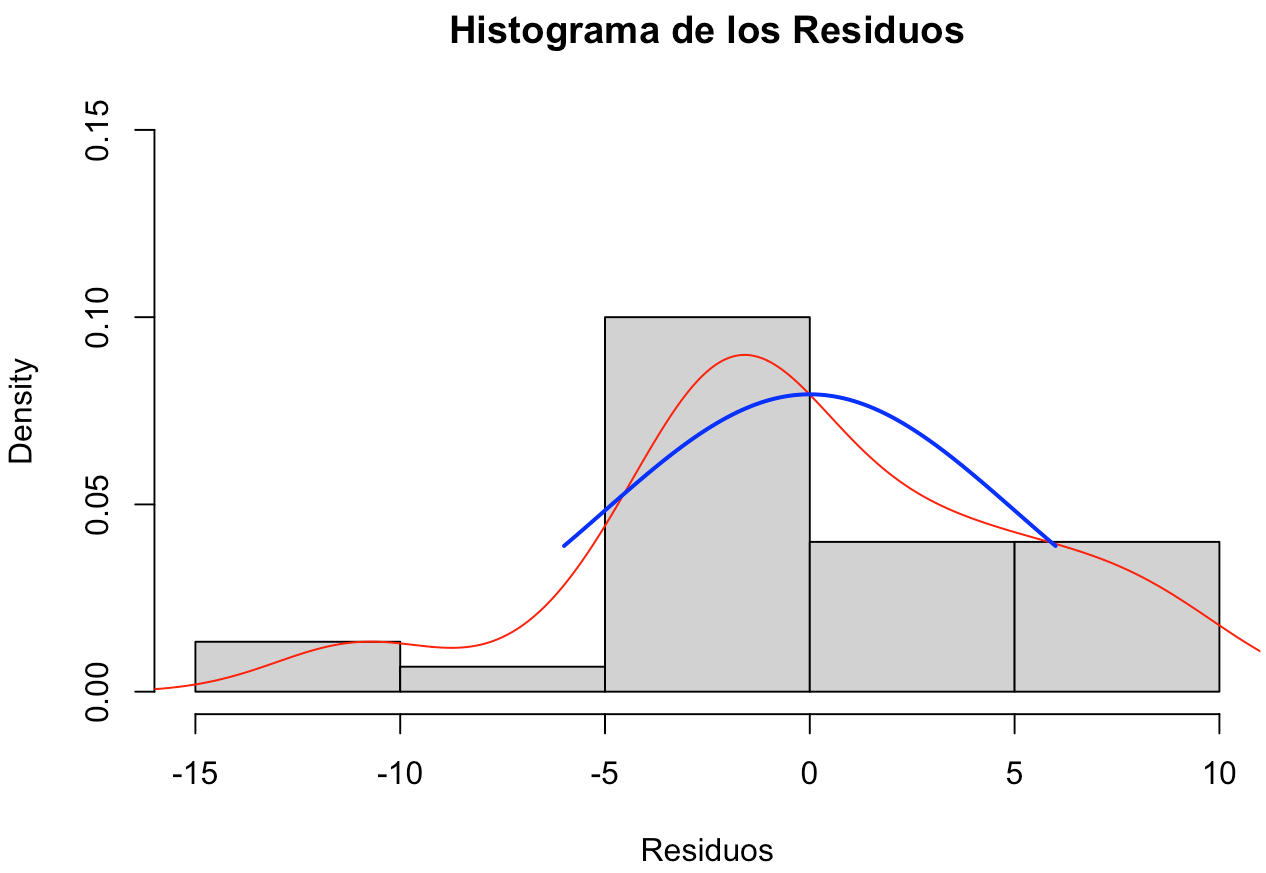
```
library(nortest)

ad.test(modelo_mixto$residuals)
```

```
##
##  Anderson-Darling normality test
##
## data:  modelo_mixto$residuals
## A = 0.41149, p-value = 0.3204
```

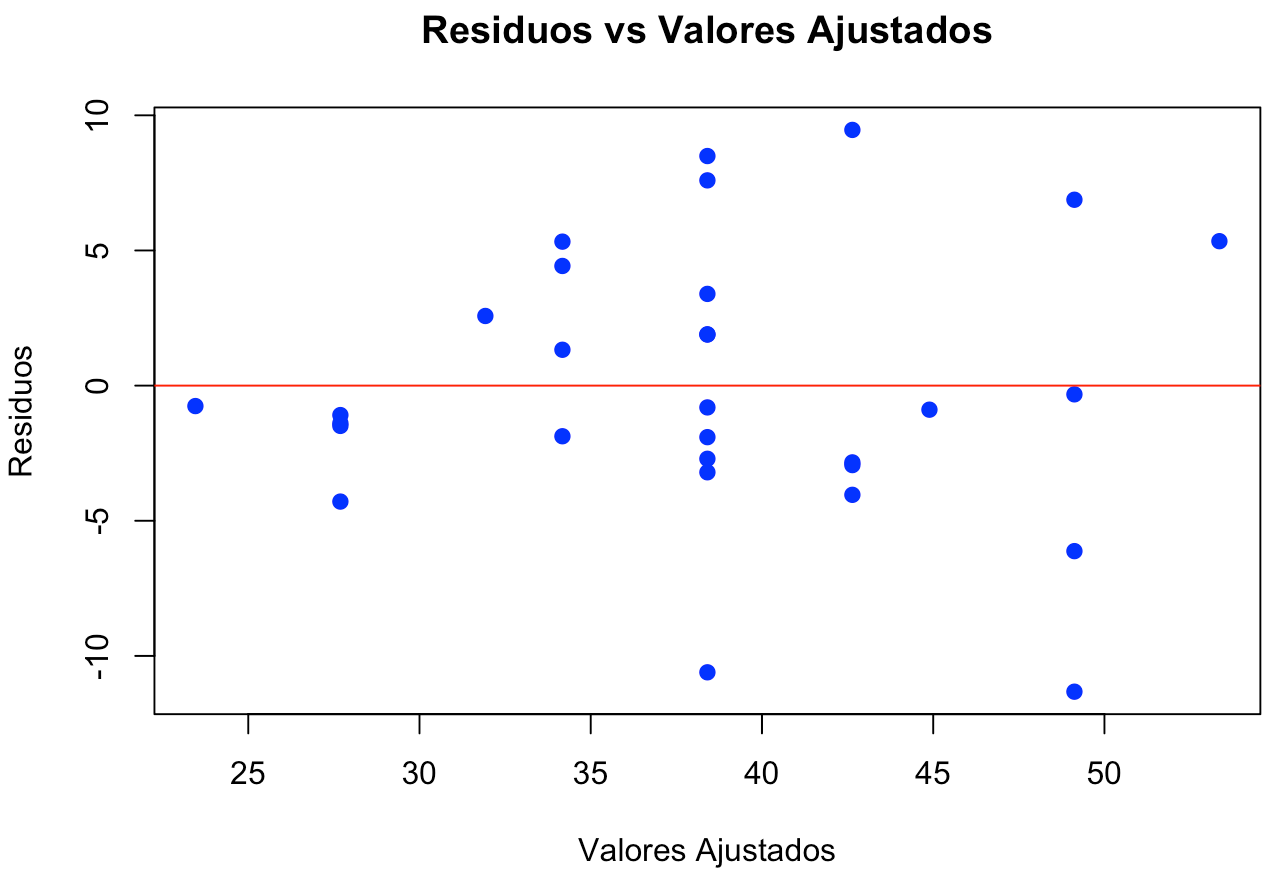
```
qqnorm(modelo_mixto$residuals, main = "Gráfico Q-Q de los Residuos")
qqline(modelo_mixto$residuals, col = "red")
```




```
hist(modelo_mixto$residuals, freq = FALSE, ylim = c(0, 0.15), main = "Histograma de los
Residuos", xlab = "Residuos")
lines(density(modelo_mixto$residuals), col = "red")
curve(dnorm(x, mean = mean(modelo_mixto$residuals), sd = sd(modelo_mixto$residuals)), fr
om = -6, to = 6, add = TRUE, col = "blue", lwd = 2)
```




```
plot(predict(modelo_forward), residuals(modelo_forward), main = "Residuos vs Valores Aju
stados", xlab = "Valores Ajustados", ylab = "Residuos", pch = 19, col = "blue")
abline(h = 0, col = "red")
```




```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
# Homocedasticidad
bptest(modelo_mixto)  # Prueba de Breusch-Pagan
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo_mixto
## BP = 4.0043, df = 2, p-value = 0.135
```

```
# Independencia
dwtest(modelo_mixto)  # Prueba de Durbin-Watson
```

```
##
##  Durbin-Watson test
##
## data:  modelo_mixto
## DW = 2.3511, p-value = 0.8267
## alternative hypothesis: true autocorrelation is greater than 0
```

```
bgtest(modelo_mixto) # Prueba de Breusch-Godfrey
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  modelo_mixto
## LM test = 1.1371, df = 1, p-value = 0.2863
```

```
# Linealidad
resettest(modelo_mixto) # Prueba de RESET de Ramsey
```

```
##
##  RESET test
##
## data:  modelo_mixto
## RESET = 0.79035, df1 = 2, df2 = 25, p-value = 0.4647
```

- No multicolinealidad de X_i

```
library(car)
```

```
## Loading required package: carData
```

```
vif(modelo_mixto)
```

```
##    Potencia Temperatura
##           1           1
```

```
cor(df[, c("Potencia", "Temperatura")])
```

```
##             Potencia Temperatura
## Potencia           1           0
## Temperatura        0           1
```

4. Emite conclusiones sobre el modelo final encontrado e interpreta en el contexto del problema el efecto de las variables predictoras en la variable respuesta.

En general se realizó un análisis para validar cual de los tres modelos (mixto, forward, backward) es mejor. Y cuál de estos predice la variable dependiente “Resistencia”, como se mostró después de aplicar cada modelo, los tres llegaron a la misma conclusión que el mejor modelo es “Resistencia ~ Potencia + Temperatura”, por supuesto cada modelo tuvo su forma especifica de como es que llegaron a esta conclusión. Esto nos indica que estas dos variables predictoras (Potencia ,Temperatura) son las mejores para ayudar a predecir Resistencia, ya que estas resultaron ser las más significativas para explicar la variabilidad de Resistencia.

Ahora al evaluar la **homocedasticidad** usando la prueba de Breusch-Pagan, tenemos las siguientes hipótesis:

- Hipótesis nula (H_0): Los residuos del modelo son homocedásticos (la varianza de los residuos es constante a lo largo de todas las observaciones).
- Hipótesis alternativa (H_1): Los residuos presentan heterocedasticidad (la varianza de los residuos no es constante).

Dado que en nuestro caso el p-valor = 0.135 es mayor que 0.05, deducimos que no hay suficiente evidencia para rechazar la hipótesis nula de homocedasticidad. Lo que significa que no se detecta heterocedasticidad en los residuos del modelo, por lo que se puede asumir que la varianza de los errores es constante.

El modelo cumple con el supuesto de homocedasticidad, lo que es una bueno. La prueba no encontró ninguna evidencia de que la varianza de los residuos cambie a lo largo de los valores predichos, lo cual valida aún más el modelo.

Ahora para la **independecia** se realizaron las siguientes pruebas:

- Prueba de Durbin-Watson
- Prueba de Breusch-Godfrey

Estás pruebas nos ayudan a detectar la autocorrelación de los residuos en un modelo de regresión, lo que buscamos es que no haya autocorrelación en nuestros residuos.

Y eso mismo nos dicen ambas pruebas, en la prueba de Durbin-Watson, el estadístico DW = 2.35 y el p-valor = 0.8267 nos dicen que los residuos no tienen autocorrelación significativa, ya que un valor cercano a 2 indica independencia de los errores.(Puede variar entre 0 y 4, 2 seria el neutro en este caso) Igual, de manera similar, la prueba de Breusch-Godfrey también mostró un p-valor = 0.2863, lo que confirma que no hay autocorrelación de orden superior en los residuos. Con ambos resultados, podemos validar el supuesto de independencia de los errores en el modelo.

Ahora explicando la **linealidad**:

Se realizó la prueba de RESET de Ramsey, lo que nos muestra un p-valor = 0.4647, esto nos dice que no hay suficiente evidencia para rechazar la hipótesis nula de que el modelo está correctamente especificado. Esto significa que la relación entre las variables predictoras y la variable dependiente es lineal, y no es necesario agregar términos no lineales (como cuadráticos o cúbicos) para mejorar el ajuste del modelo.

VIF y Matriz de correlación

El VIF mide cuánto la varianza de un coeficiente de regresión está “inflada” debido a la colinealidad con otras variables predictoras. Un VIF = 1 significa que no hay correlación entre una variable y las demás.

En nuestro casol os resultados que nos da para VIF muestran que tanto Potencia como Temperatura tienen un VIF = 1. Esto es lo mejor que se puede obtener, ya que significa que no existe ninguna colinealidad entre las variables.

Con esto, podemos tener una pequeña idea de como se vería la matriz de correlación.

Correlación 0 entre Potencia y Temperatura significa que no hay relación lineal entre estas variables. Esto confirma que no hay problemas de multicolinealidad.

Y con esto hemos concluido que el mejor modelo es “Resistencia ~ Potencia + Temperatura”, donde se observa que solo necesitamos de dos variables, lo cual hace el modelo simple, pero efectivo.