

Actividad 2: Regresión Múltiple

Daniela Jiménez Téllez

2024-09-17

Problema

En la base de datos Al corte se describe un experimento realizado para evaluar el impacto de las variables: fuerza, potencia, temperatura y tiempo sobre la resistencia al corte. Indica cuál es la mejor relación entre estas variables que describen la resistencia al corte.

Importación de librerías

```
library(lmtest)
```

```
## Cargando paquete requerido: zoo
```

```
##  
## Adjuntando el paquete: 'zoo'
```

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

```
library(car)
```

```
## Cargando paquete requerido: carData
```

Importación de datos

```
datos <- read.csv("AlCorte.csv")
```

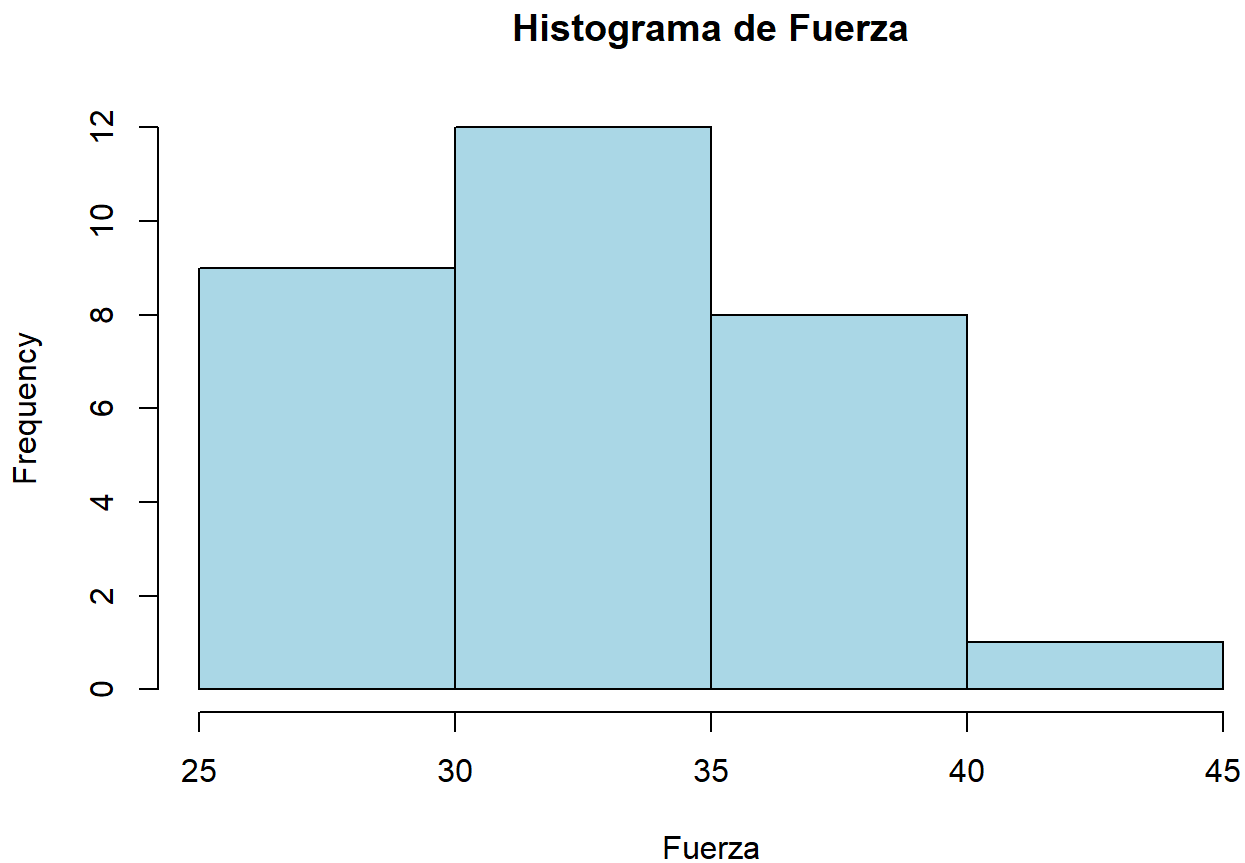
1. Haz un análisis descriptivo de los datos: medidas principales y gráficos

```
# Resumen de medidas  
  
print(summary(datos))
```

##	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

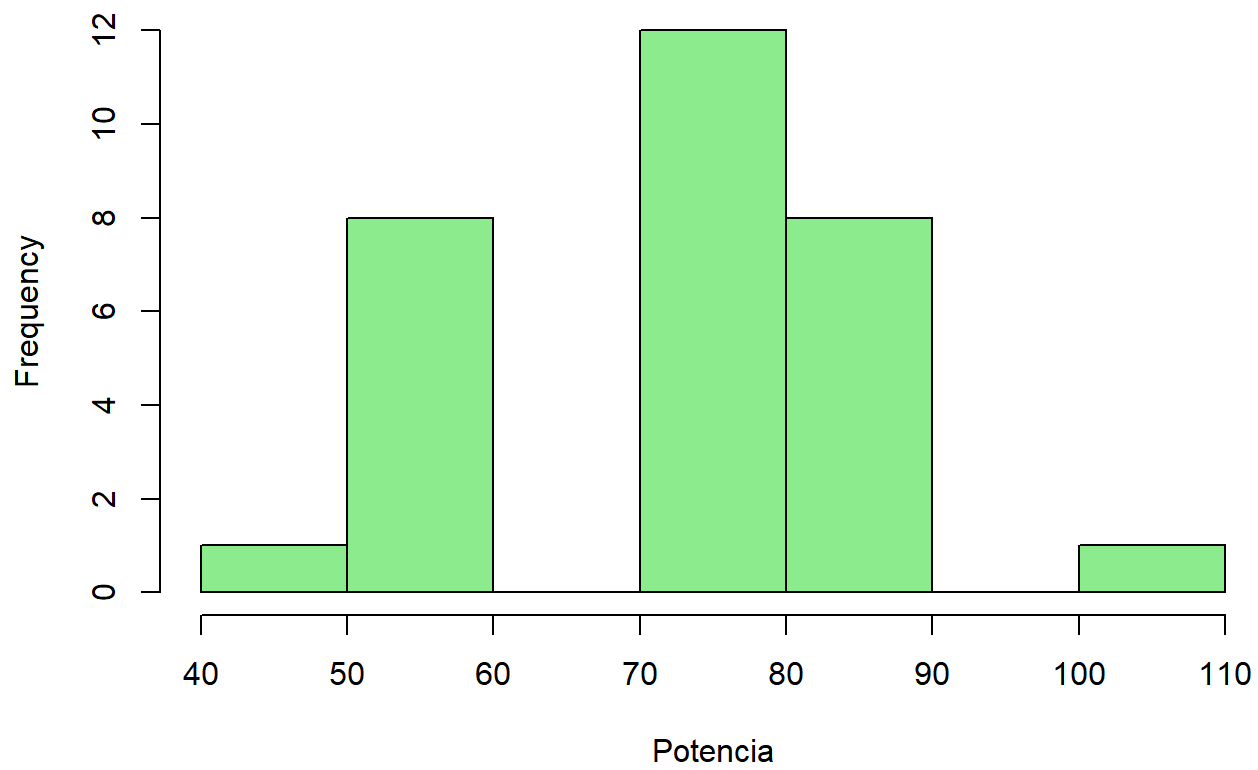
Histogramas

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



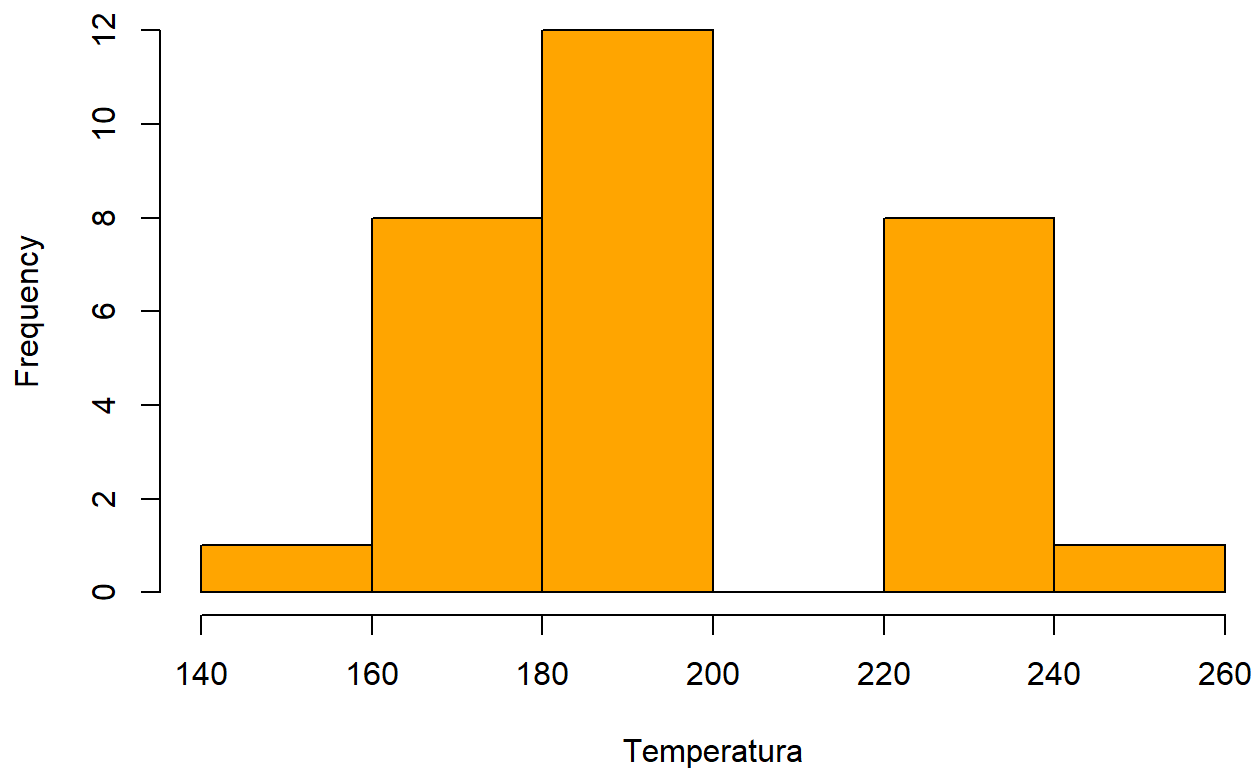
```
hist(datos$Potencia, main = "Histograma de Potencia", xlab = "Potencia", col = "lightgreen")
```

Histograma de Potencia



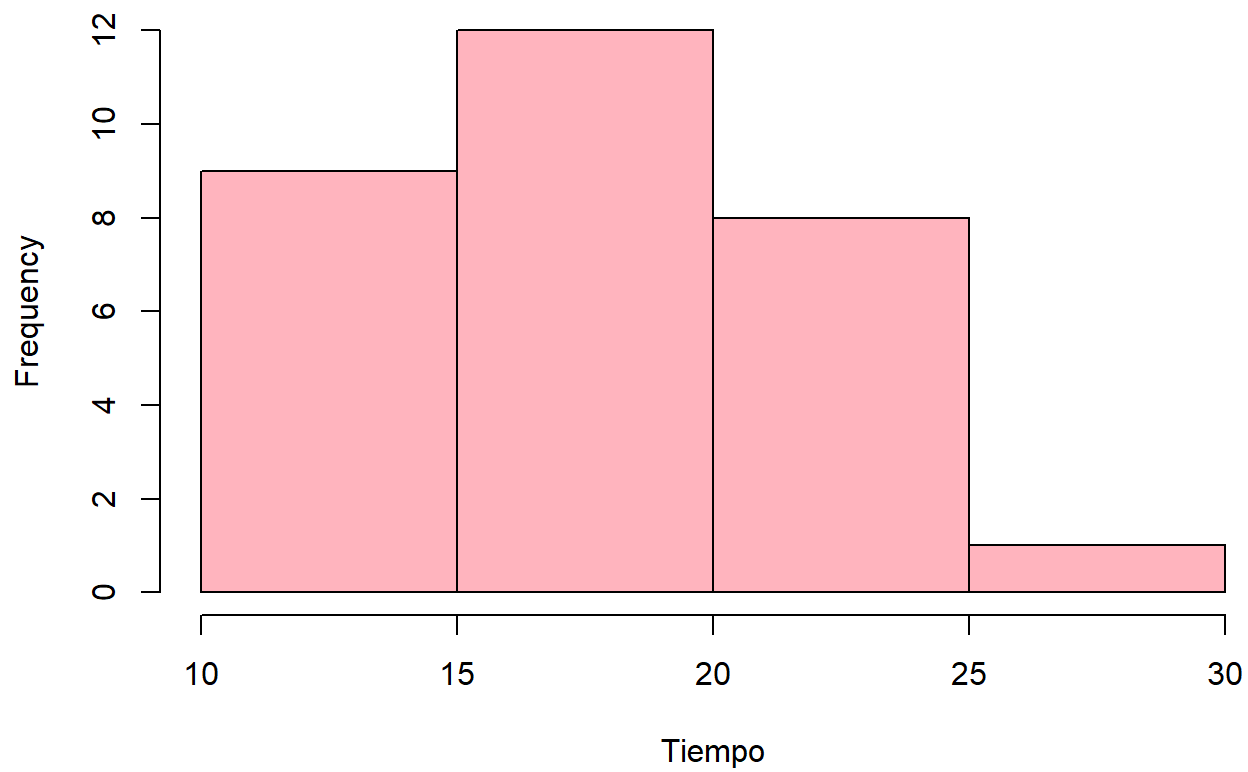
```
hist(datos$Temperatura, main = "Histograma de Temperatura", xlab = "Temperatura", col = "orange")
```

Histograma de Temperatura



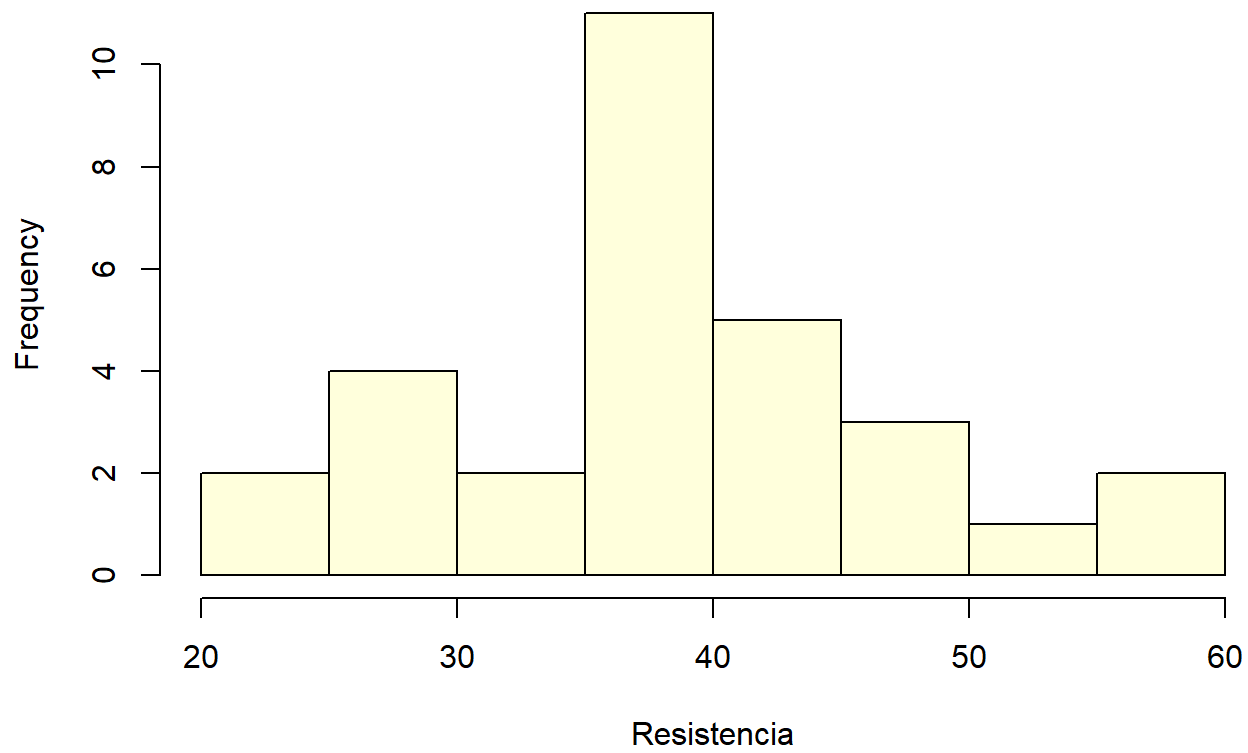
```
hist(datos$Tiempo, main = "Histograma de Tiempo", xlab = "Tiempo", col = "lightpink")
```

Histograma de Tiempo



```
hist(datos$Resistencia, main = "Histograma de Resistencia", xlab = "Resistencia", col = "lightyellow")
```

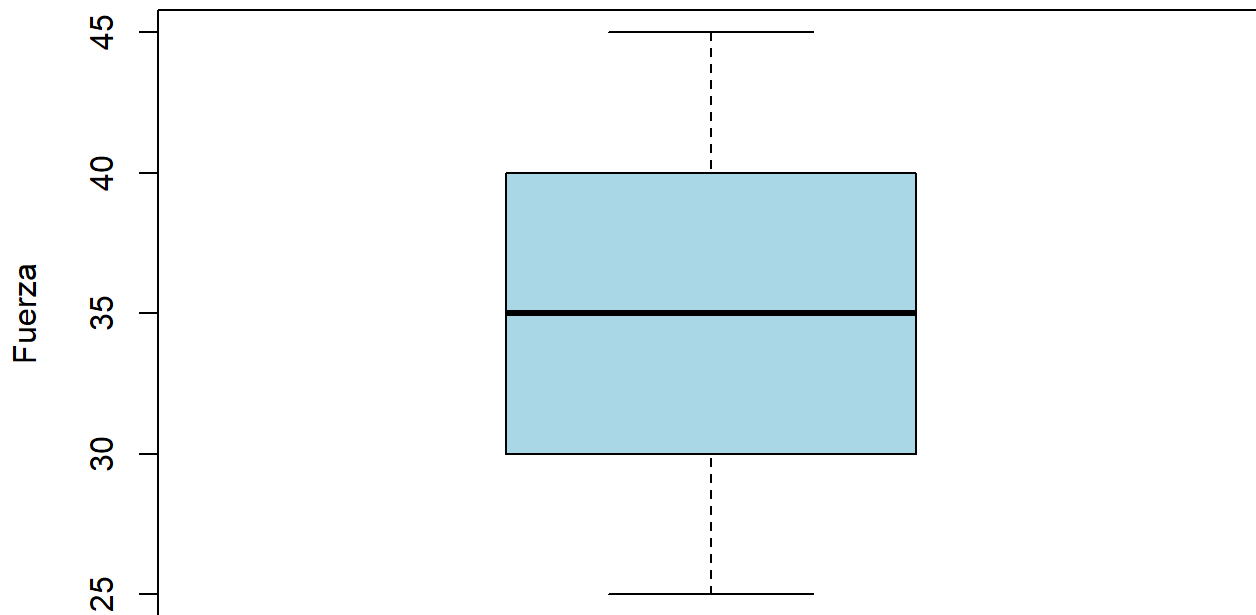
Histograma de Resistencia



```
# Boxplots
```

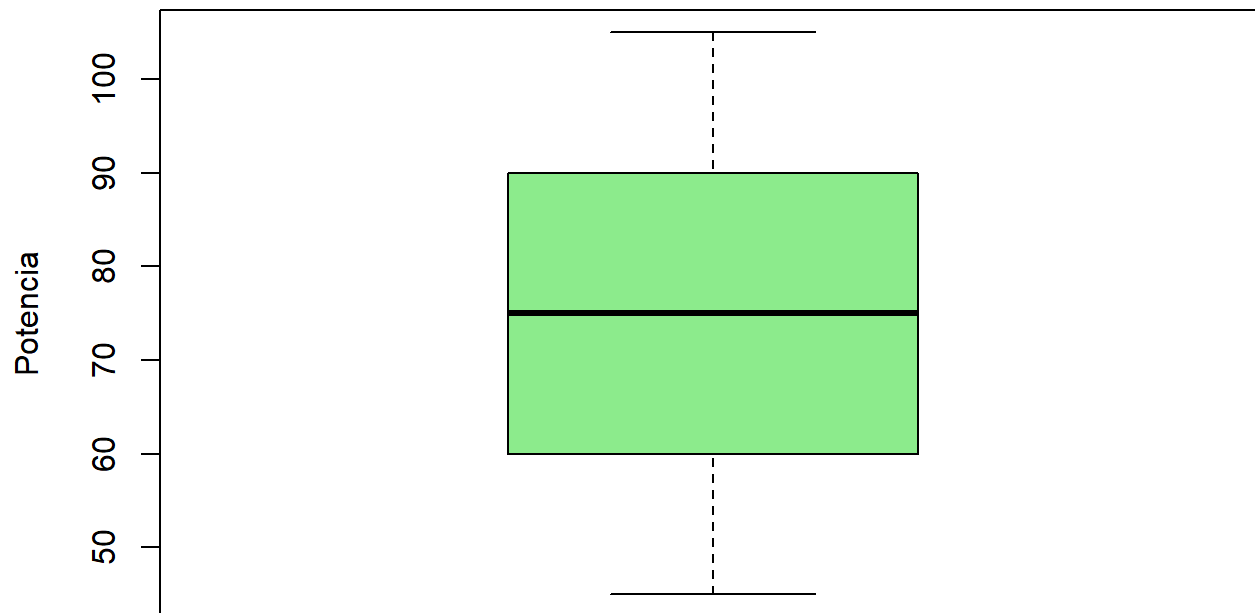
```
boxplot(datos$Fuerza, main = "Boxplot de Fuerza", ylab = "Fuerza", col = "lightblue")
```

Boxplot de Fuerza



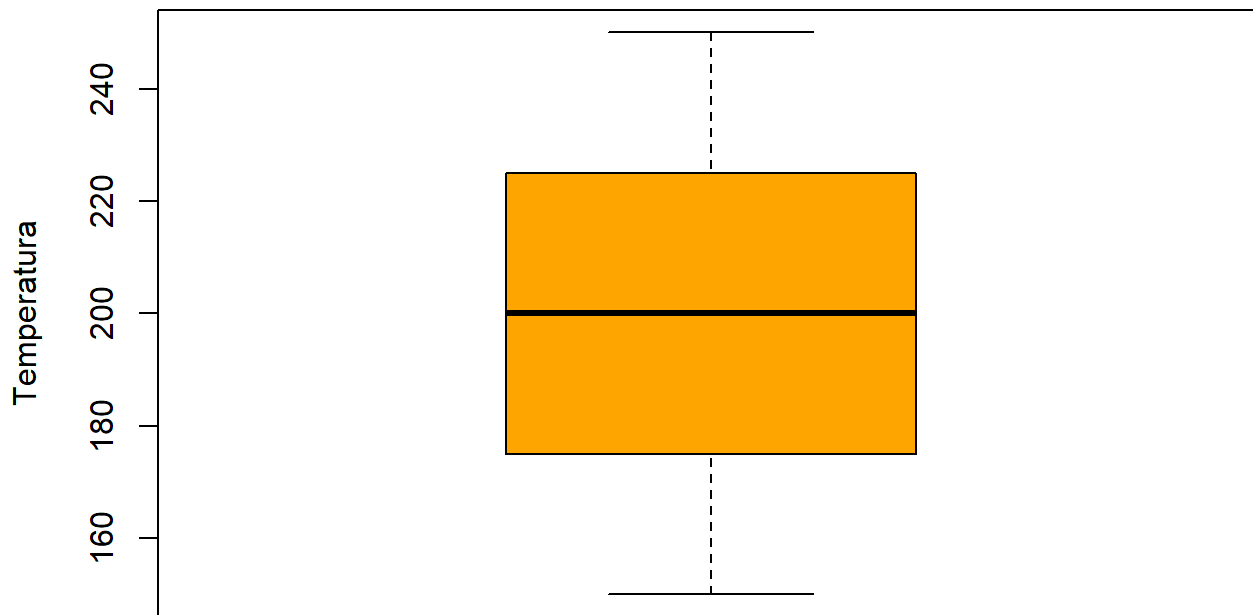
```
boxplot(datos$Potencia, main = "Boxplot de Potencia", ylab = "Potencia", col = "lightgreen")
```

Boxplot de Potencia



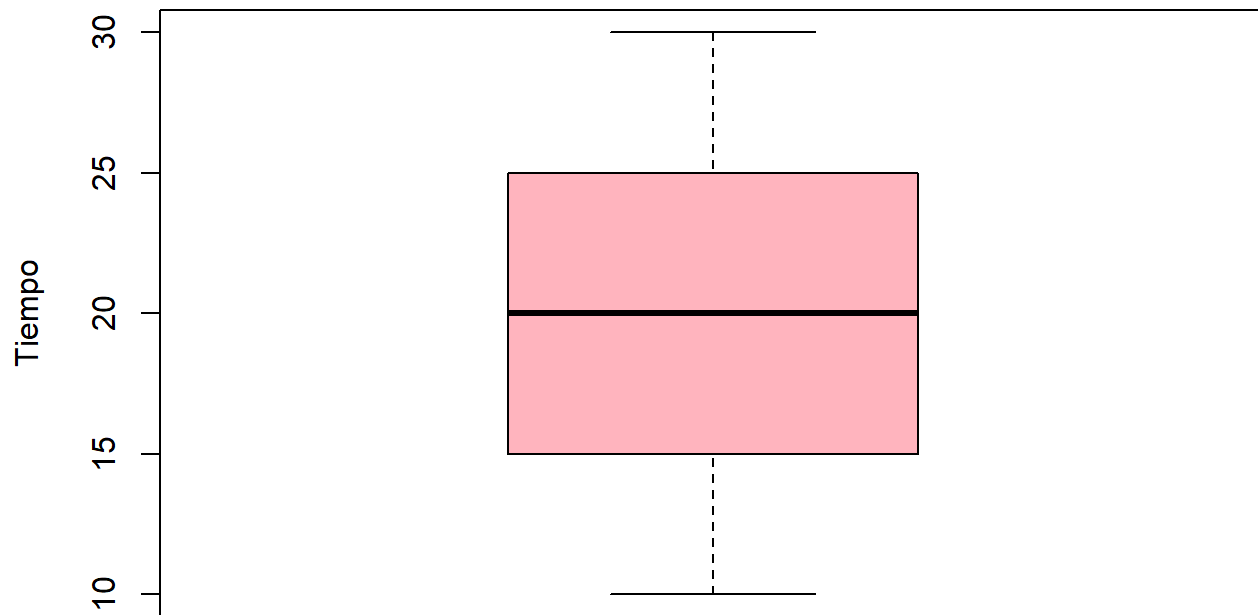
```
boxplot(datos$Temperatura, main = "Boxplot de Temperatura", ylab = "Temperatura", col = "orange")
```


Boxplot de Temperatura



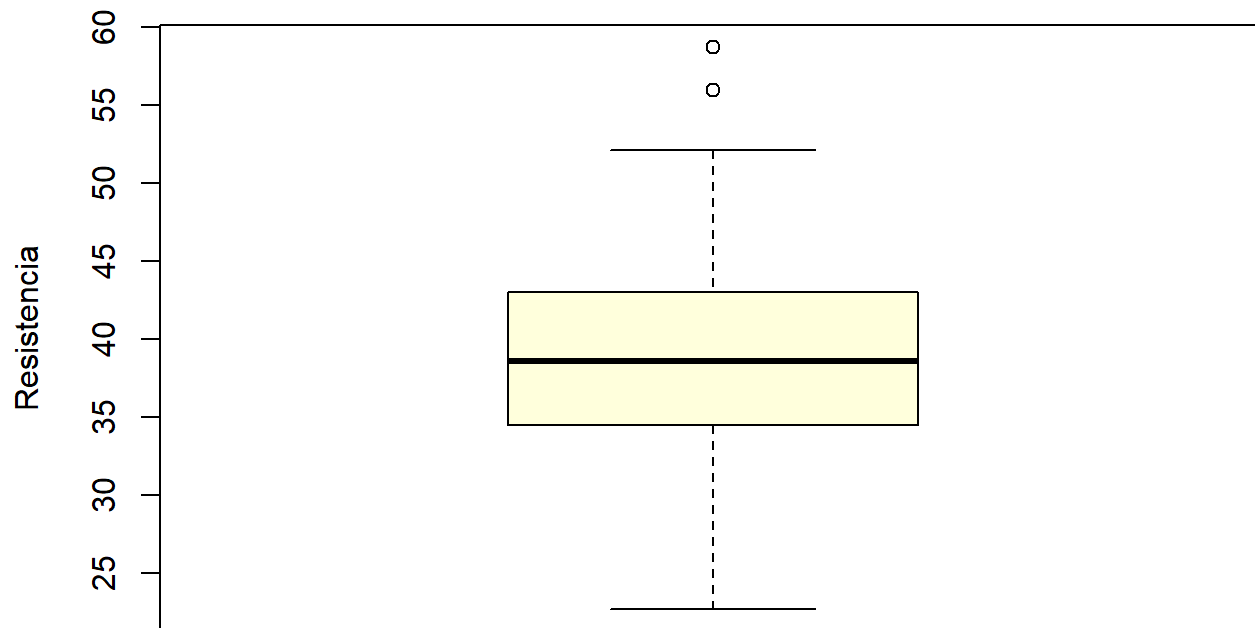
```
boxplot(datos$Tiempo, main = "Boxplot de Tiempo", ylab = "Tiempo", col = "lightpink")
```

Boxplot de Tiempo



```
boxplot(datos$Resistencia, main = "Boxplot de Resistencia", ylab = "Resistencia", col = "lightyellow")
```

Boxplot de Resistencia



2. Encuentra el mejor modelo de regresión que explique la variable Resistencia. Analiza el modelo basándote en la significancia:

```
modeloCompleto <- lm(Resistencia ~ ., data = datos)
modeloNulo <- lm(Resistencia ~ 1, data = datos)

# Modelo Mixto

pasosMixto <- step(modeloCompleto, 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   1341.02 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
```

```
cat("\n\n")
```

```
# Modelo Forward
```

```
pasosForward <- step(modeloNulo, scope = list(lower = modeloNulo, upper = modeloCompleto), 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
```

```
cat("\n\n")
```

```
# Modelo Backward
```

```
pasosBackward <- step(modeloCompleto, 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   1341.02 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
```

Observando los resultados anteriores se puede observar que el mejor modelo es el que incluye Potencia y Temperatura. Se comprobará con las siguientes pruebas:

```
# Modelo
```

```
modelo <- lm(Resistencia ~ Potencia + Temperatura, data = datos)
```

- Economía de las variables
- Significación global (Prueba para el modelo)
- Significación individual (Prueba para cada β_i)
- Variación explicada por el modelo

```
# Significancia global

resumen <- summary(modelo)

# Significancia individual

sigIndividual <- confint(modelo)

# Variación

r2 <- resumen$adj.r.squared

# Prints

cat("Summary: \n\n")
```

```
## Summary:
```

```
print(resumen)
```

```
##
## Call:
## lm(formula = Resistencia ~ Potencia + Temperatura, data = datos)
##
## 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("\n\n")
```

```
cat("Significancia Individual: \n\n")
```

```
## Significancia Individual:
```

```
print(sigIndividual)
```

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

```
cat("\n\n")
```

```
cat("Variación del modelo: \n\n")
```

```
## Variación del modelo:
```

```
cat("R² ajustado =", r2, "\n")
```

```
## R² ajustado = 0.6618581
```

3. Analiza la validez del modelo encontrado:

- Análisis de residuos (homocedasticidad, independencia, etc)
- No multicolinealidad de X_i

```
# Media cero

mediaCero <- t.test(resid(modelo))

# Normalidad

norm <- shapiro.test(resid(modelo))

# Homocedasticidad

hom1 <- bptest(modelo)
hom2 <- gqtest(modelo)

# Independencia

ind1 <- dwtest(modelo)
ind2 <- bgtest(modelo)

# No multicolinealidad

multi <- vif(modelo)

# Prints

cat("Media Cero en los Residuos del Modelo:\n\n")
```

```
## Media Cero en los Residuos del Modelo:
```



```
print(mediaCero)
```

```
##  
## One Sample t-test  
##  
## data: resid(modelo)  
## t = 8.8667e-17, df = 29, p-value = 1  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## -1.876076 1.876076  
## sample estimates:  
## mean of x  
## 8.133323e-17
```

```
cat("\n\n")
```

```
cat("Prueba de Normalidad (Shapiro-Wilk) para los Residuos del Modelo:\n")
```

```
## Prueba de Normalidad (Shapiro-Wilk) para los Residuos del Modelo:
```

```
print(norm)
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: resid(modelo)  
## W = 0.96588, p-value = 0.4333
```

```
cat("\n\n")
```

```
cat("Prueba de Homocedasticidad Breusch-Pagan para el Modelo:\n")
```

```
## Prueba de Homocedasticidad Breusch-Pagan para el Modelo:
```

```
print(hom1)
```

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

```
cat("\n")
```

```
cat("Prueba de Homocedasticidad Goldfeld-Quandt para el Modelo:\n")
```

```
## Prueba de Homocedasticidad Goldfeld-Quandt para el Modelo:
```

```
print(hom2)
```

```
##  
## Goldfeld-Quandt test  
##  
## data: modelo  
## GQ = 0.9753, df1 = 12, df2 = 12, p-value = 0.5169  
## alternative hypothesis: variance increases from segment 1 to 2
```

```
cat("\n\n")
```

```
cat("Prueba de Independencia Durbin-Watson para el Modelo:\n")
```

```
## Prueba de Independencia Durbin-Watson para el Modelo:
```

```
print(ind1)
```

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

```
cat("\n")
```

```
cat("Prueba de Independencia Breusch-Godfrey para el Modelo:\n")
```

```
## Prueba de Independencia Breusch-Godfrey para el Modelo:
```

```
print(ind2)
```

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

```
cat("\n\n")
```

```
cat("Verificación de Multicolinealidad (VIF) para el Modelo:\n")
```

```
## Verificación de Multicolinealidad (VIF) para el Modelo:
```

```
print(multi)
```

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

```
cat("\n\n")
```

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

Habiendo hecho todos los análisis anteriores se puede concluir que el modelo que mejor explica la Resistencia es el que incluye las variables Potencia y Temperatura. Esto se puede observar en los tres enfoques: mixto, forward y backward. Es eficiente debido a su bajo AIC, además de que tiene un R^2 ajustado de 0.66, lo que es relativamente alto en comparación a los otros. Igualmente, en las pruebas de residuos podemos observar que los datos se comportan como una distribución normal, y no existe multicolinealidad entre las variables.

Finalmente, se puede decir que la Potencia tiene un coeficiente de 0.498, lo que nos dice que cada que aumenta la Potencia, se espera que la resistencia aumente 0.498 unidades. Por otro lado, en el caso de la temperatura, se ve que el coeficiente estimado es de 0.129, lo que implica que cada que suba la temperatura, la resistencia aumentará 0.129 unidades. Esto en el contexto del problema nos dice que ambas variables son importantes en el aumento de la Resistencia.