Momento de Retroalimentacion 2 - Modulo 1

Facundo Vecchi A01283666

8 de septiembre de 2022

```
library(dplyr);
```

Importacion de datos y librerias

```
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(modeest);
library(Hmisc);
```

```
## Loading required package: lattice
## Loading required package: survival
## Loading required package: Formula
## Loading required package: ggplot2
##
## Attaching package: 'Hmisc'
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
library(reshape2);
library(ggplot2);
library(car);
```

```
## Loading required package: carData
##
## Attaching package: 'car'
## The following object is masked from 'package:dplyr':
##
##     recode
```

```r
data <- read.csv("mercurio.csv")
colnames(data) <- c("Id", "Lago", "Alcalinidad", "PH", "Calcio", "Clorofila", "con_med_mercurio",
                    "num_peces", "min_con_mercurio", "max_con_mercurio", "prom_mercurio_pez", "edad")
```

Descripcion de datos

```r
data_temp <- subset(data, select = -Id)

for (col in names(data_temp)) {
  if (class(data_temp[, col]) == "integer" | class(data_temp[, col]) == "numeric") {
    c <- nchar(col)
    cl <- 0
    cr <- 0
    if (c %% 2 != 0) {
      cl <- c / 2
      cr <- c / 2
    } else {
      cl <- c / 2
      cr <- c / 2 + 1
    }

    cat(strrep('-', 30 - cl), col, strrep('-', 30 - cr), "\n")
    cat("Promedio: ", mean(data_temp[, col]), " ",
        "Mediana: ", median(data_temp[, col]), " ",
        "Moda: ", mfv(data_temp[, col]), "\n")
    cat("Desviacion estandar: ", sd(data_temp[, col]),
        " ", "Varianza: ", var(data_temp[, col]), "\n")
    cat("Minimo: ", min(data_temp[, col]), " ",
        "Maximo: ", max(data_temp[, col]), "\n")
    cat("\n")
  }
}
```

```
## ----------------------- Alcalinidad ------------------------
## Promedio:  37.53019   Mediana:  19.6   Moda:  17.3 25.4
## Desviacion estandar:  38.20353   Varianza:  1459.509
## Minimo:  1.2   Maximo:  128
##
## ---------------------------- PH ----------------------------
## Promedio:  6.590566   Mediana:  6.8   Moda:  5.8 6.9
## Desviacion estandar:  1.288449   Varianza:  1.660102
## Minimo:  3.6   Maximo:  9.1
##
## -------------------------- Calcio --------------------------
## Promedio:  22.20189   Mediana:  12.6   Moda:  3 3.3 5.2 6.3 20.5
## Desviacion estandar:  24.93257   Varianza:  621.6333
## Minimo:  1.1   Maximo:  90.7
##
## ------------------------ Clorofila -------------------------
## Promedio:  23.11698   Mediana:  12.8   Moda:  1.6 3.2 9.6
## Desviacion estandar:  30.81632   Varianza:  949.6457
## Minimo:  0.7   Maximo:  152.4
##
```

```
## ---------------------- con_med_mercurio ----------------------
## Promedio:  0.5271698   Mediana:  0.48   Moda:  0.34
## Desviacion estandar:  0.3410356   Varianza:  0.1163053
## Minimo:  0.04   Maximo:  1.33
##
## ------------------------- num_peces -------------------------
## Promedio:  13.0566   Mediana:  12   Moda:  12
## Desviacion estandar:  8.560677   Varianza:  73.2852
## Minimo:  4   Maximo:  44
##
## ---------------------- min_con_mercurio ----------------------
## Promedio:  0.2798113   Mediana:  0.25   Moda:  0.04
## Desviacion estandar:  0.2264058   Varianza:  0.05125958
## Minimo:  0.04   Maximo:  0.92
##
## ---------------------- max_con_mercurio ----------------------
## Promedio:  0.8745283   Mediana:  0.84   Moda:  0.06 0.26 0.4 0.48 0.69 0.84 1.4 1.5 1.9
## Desviacion estandar:  0.5220469   Varianza:  0.2725329
## Minimo:  0.06   Maximo:  2.04
##
## --------------------- prom_mercurio_pez ----------------------
## Promedio:  0.5132075   Mediana:  0.45   Moda:  0.16
## Desviacion estandar:  0.3387294   Varianza:  0.1147376
## Minimo:  0.04   Maximo:  1.53
##
## ---------------------------- edad ----------------------------
## Promedio:  0.8113208   Mediana:  1   Moda:  1
## Desviacion estandar:  0.3949977   Varianza:  0.1560232
## Minimo:  0   Maximo:  1
```

```r
for (col in names(data_temp)) {
  if (class(data_temp[, col]) == "character") {
    c <- nchar(col)
    cl <- 0
    cr <- 0
    if (c %% 2 != 0) {
      cl <- c / 2
      cr <- c / 2
    } else {
      cl <- c / 2
      cr <- c / 2 + 1
    }

    cat(strrep('-', 30 - cl), col, strrep('-', 30 - cr), "\n")
    cat("Moda: ", mfv(data_temp[, col]), "\n")
    cat("\n")
    print(table(data_temp[, col]))
    cat("\n")
  }
}
```

```
## ---------------------------- Lago ----------------------------
## Moda:  Alligator Annie Apopka Blue Cypress Brick Bryant Cherry Crescent Deer Point Dias Dorr Down Eas
##
##
```

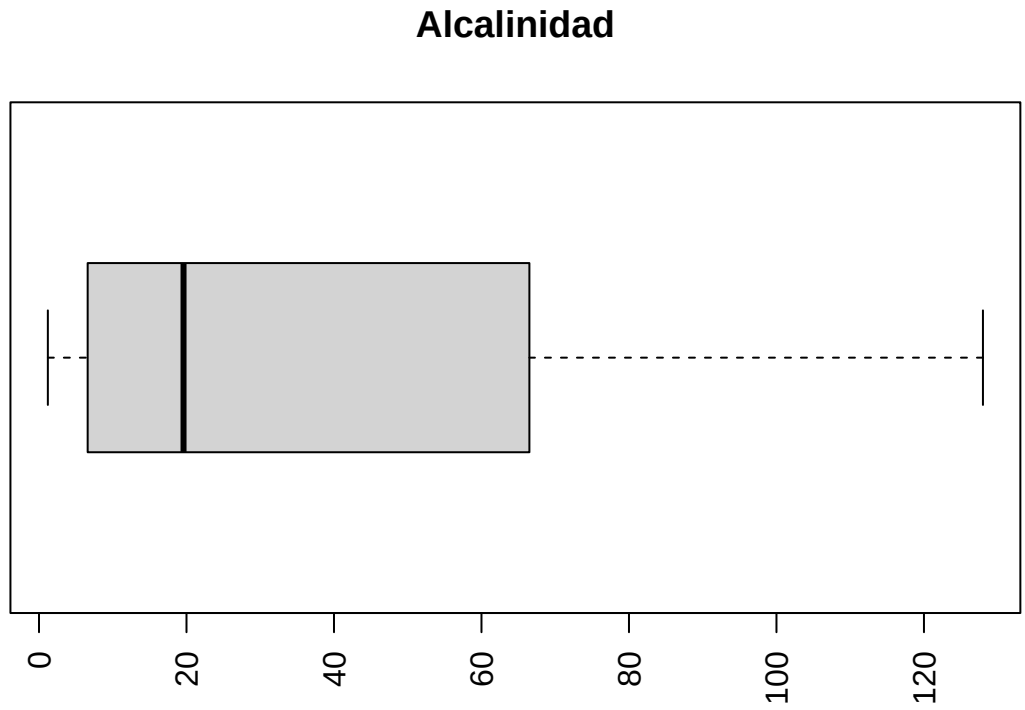
```
##         Alligator             Annie             Apopka      Blue Cypress
##                 1                 1                  1                 1
##             Brick            Bryant            Cherry          Crescent
##                 1                 1                  1                 1
##        Deer Point              Dias              Dorr              Down
##                 1                 1                  1                 1
## East Tohopekaliga             Eaton           Farm-13            George
##                 1                 1                  1                 1
##           Griffin            Harney              Hart        Hatchineha
##                 1                 1                  1                 1
##           Iamonia         Istokpoga           Jackson         Josephine
##                 1                 1                  1                 1
##          Kingsley         Kissimmee         Lochloosa            Louisa
##                 1                 1                  1                 1
##        Miccasukee          Minneola            Monroe           Newmans
##                 1                 1                  1                 1
##        Ocean Pond      Ocheese Pond        Okeechobee            Orange
##                 1                 1                  1                 1
##       Panasoffkee            Parker            Placid            Puzzle
##                 1                 1                  1                 1
##            Rodman          Rousseau           Sampson             Shipp
##                 1                 1                  1                 1
##           Talquin            Tarpon      Tohopekaliga          Trafford
##                 1                 1                  1                 1
##             Trout      Tsala Apopka              Weir          Wildcat
##                 1                 1                  1                 1
##              Yale
##                 1
```

Quartiles

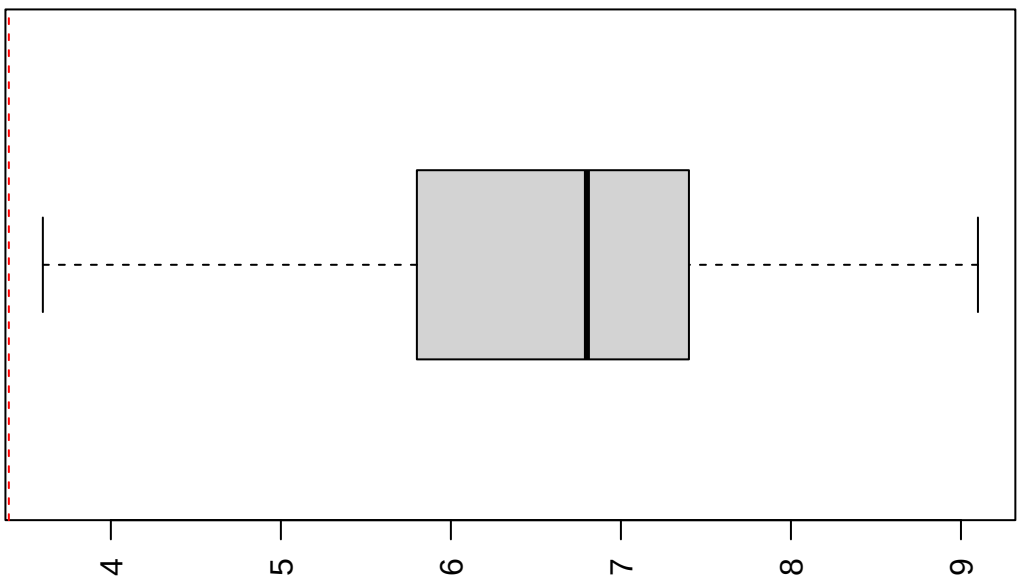
```r
for (col in names(data_temp)) {
  if (class(data_temp[, col]) == "integer" | class(data_temp[, col]) == "numeric") {
    x <- data_temp[, col]
    q <- quantile(x, c(0.25, 0.75))
    ri <- q[2] - q[1]

    c <- nchar(col)
    cl <- 0
    cr <- 0
    if (c %% 2 != 0) {
      cl <- c / 2
      cr <- c / 2
    } else {
      cl <- c / 2
      cr <- c / 2 + 1
    }

    cat(strrep('-', 30 - cl), col, strrep('-', 30 - cr), "\n")
    cat("Quartil 1: ", q[1], " ", "Quartil 3: ", q[2], "\n")
    boxplot(x, main = col, las = 2, xlab = "", ylab = "", horizontal = TRUE)
    abline(v = q[1] - 1.5 * ri, lty = 2, col = "red")
    abline(v = q[2] + 1.5 * ri, lty = 2, col = "red")
```

```
    abline(v = q[1] - 3 * ri, lty = 2, col = "blue")
    abline(v = q[2] + 3 * ri, lty = 2, col = "blue")
  }
}
```

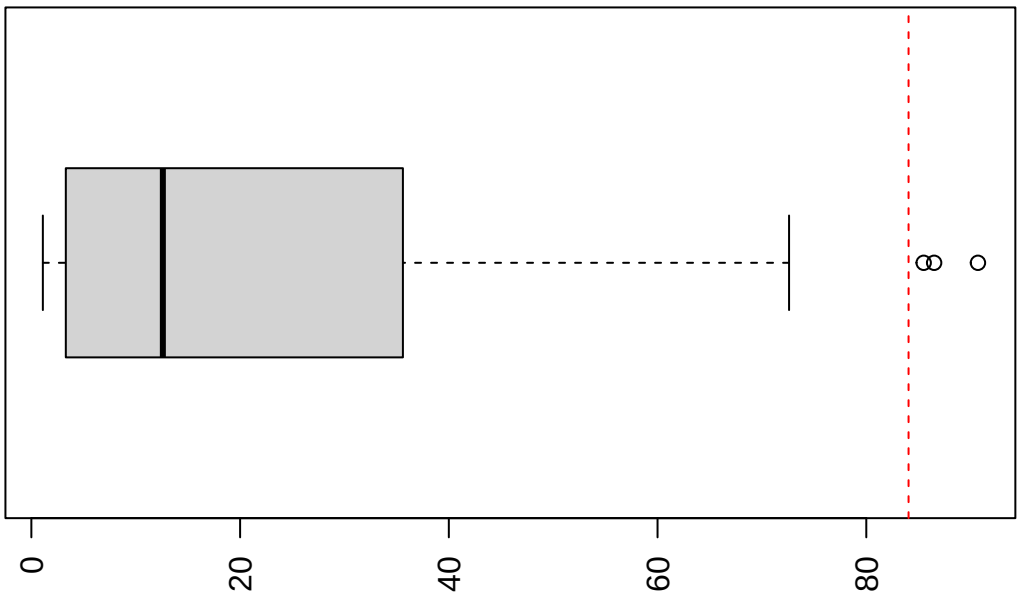
```
## ------------------------ Alcalinidad ------------------------
## Quartil 1:  6.6   Quartil 3:  66.5
```



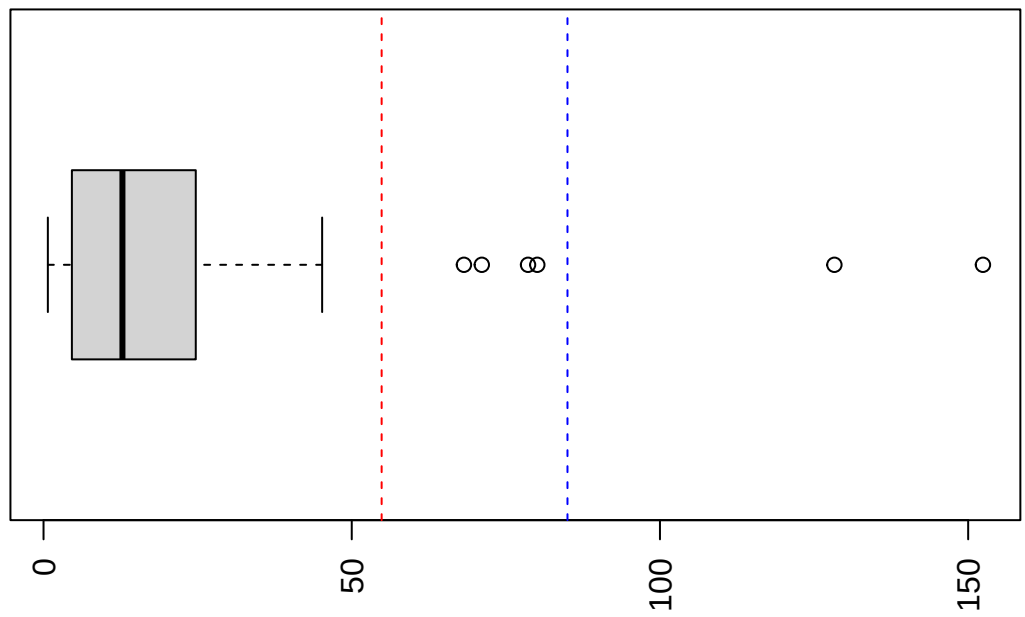

```
## ---------------------------- PH ----------------------------
## Quartil 1:  5.8   Quartil 3:  7.4
```



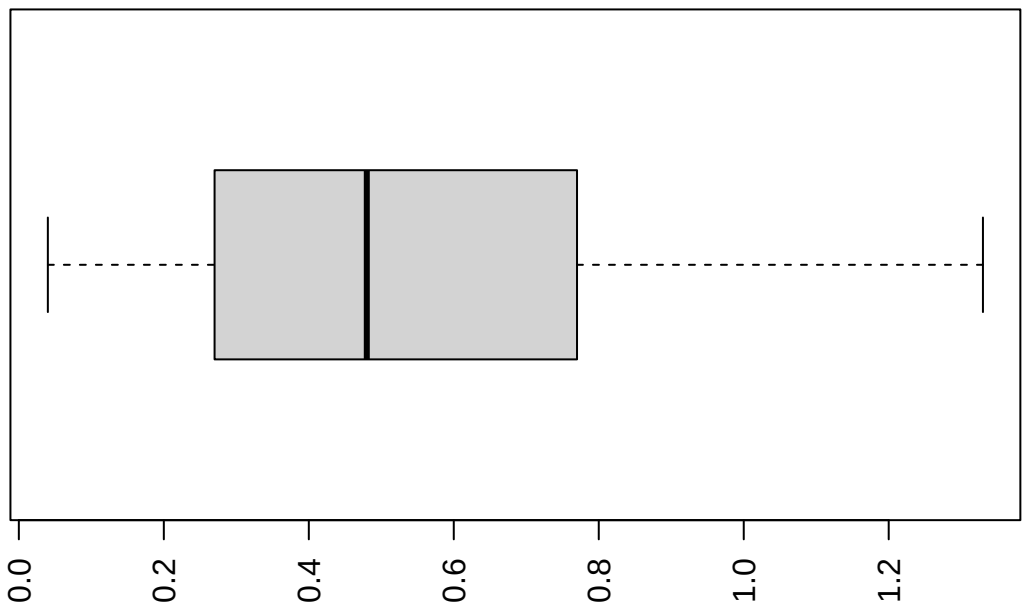

```
## --------------------------- Calcio ---------------------------
## Quartil 1:  3.3   Quartil 3:  35.6
```



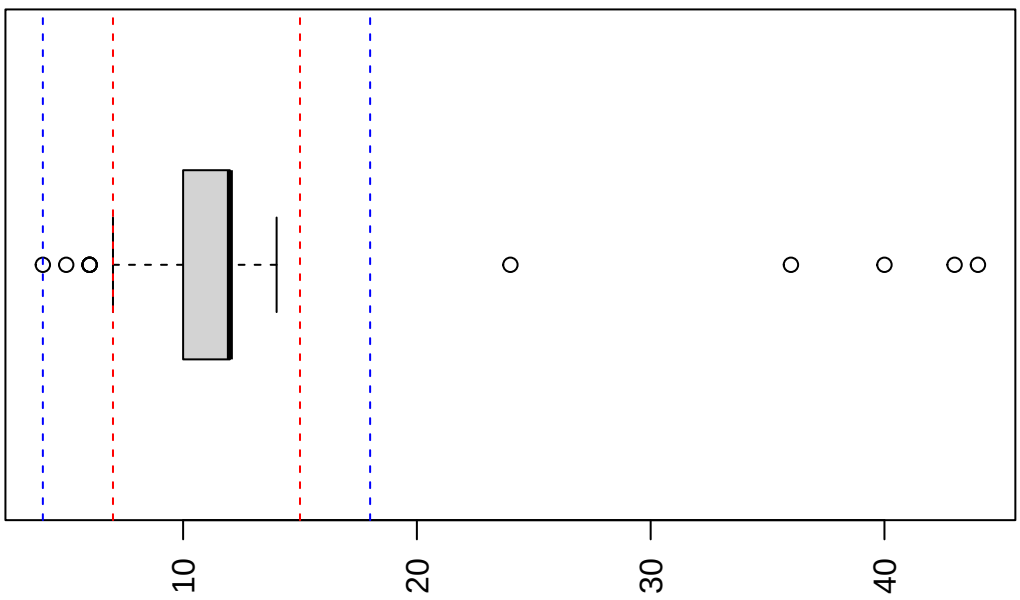

```
## ------------------------- Clorofila -------------------------
## Quartil 1:  4.6   Quartil 3:  24.7
```



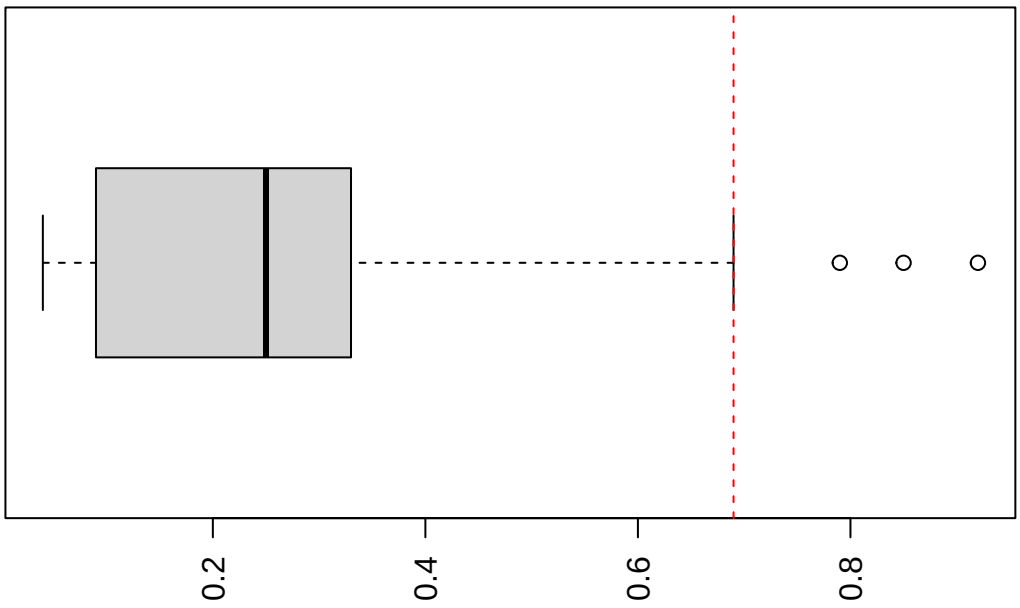

```
## ---------------------- con_med_mercurio ----------------------
## Quartil 1:  0.27   Quartil 3:  0.77
```



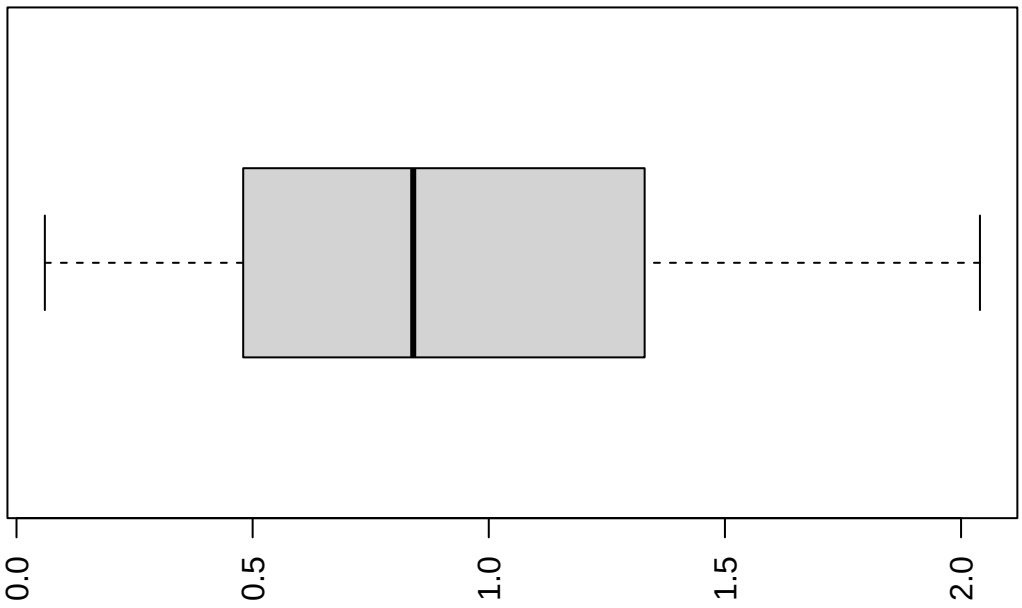

```
## ------------------------ num_peces -------------------------
## Quartil 1:  10   Quartil 3:  12
```



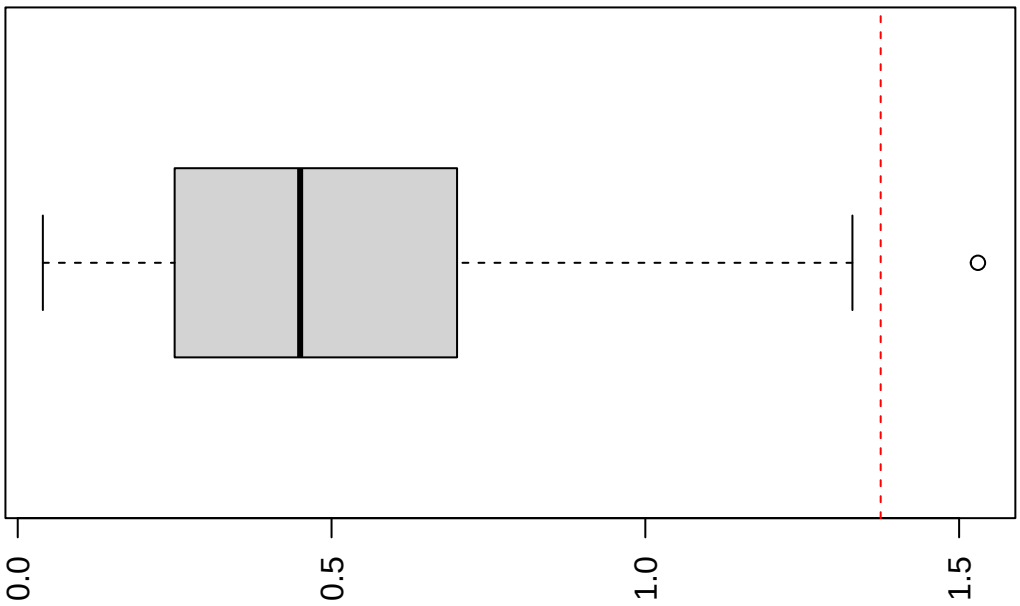

```
## ---------------------- min_con_mercurio ----------------------
## Quartil 1:  0.09   Quartil 3:  0.33
```



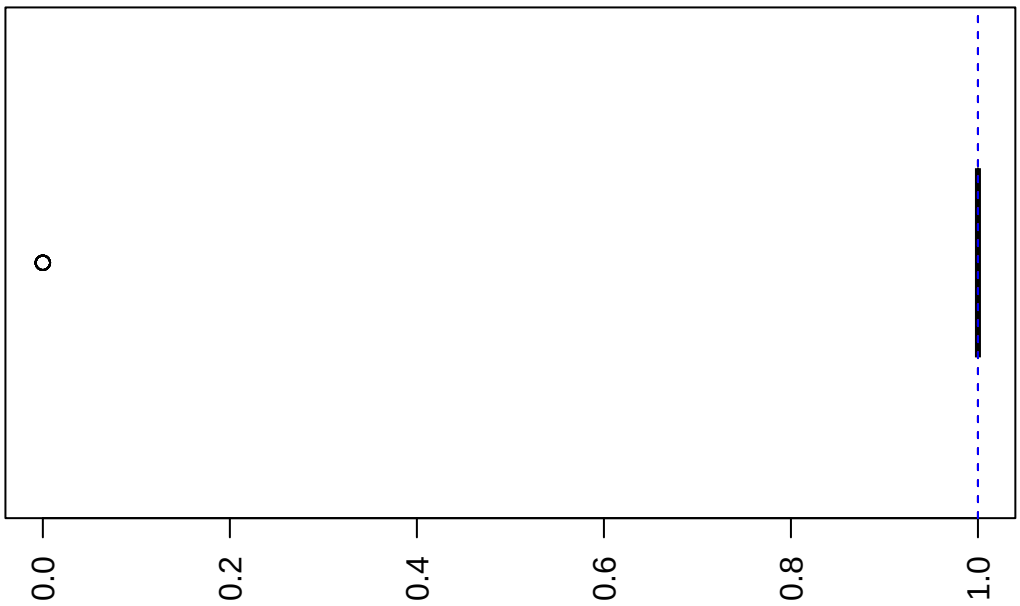

```
## ---------------------- max_con_mercurio ----------------------
## Quartil 1:  0.48   Quartil 3:  1.33
```




```
## ---------------------- prom_mercurio_pez ----------------------
## Quartil 1:  0.25   Quartil 3:  0.7
```

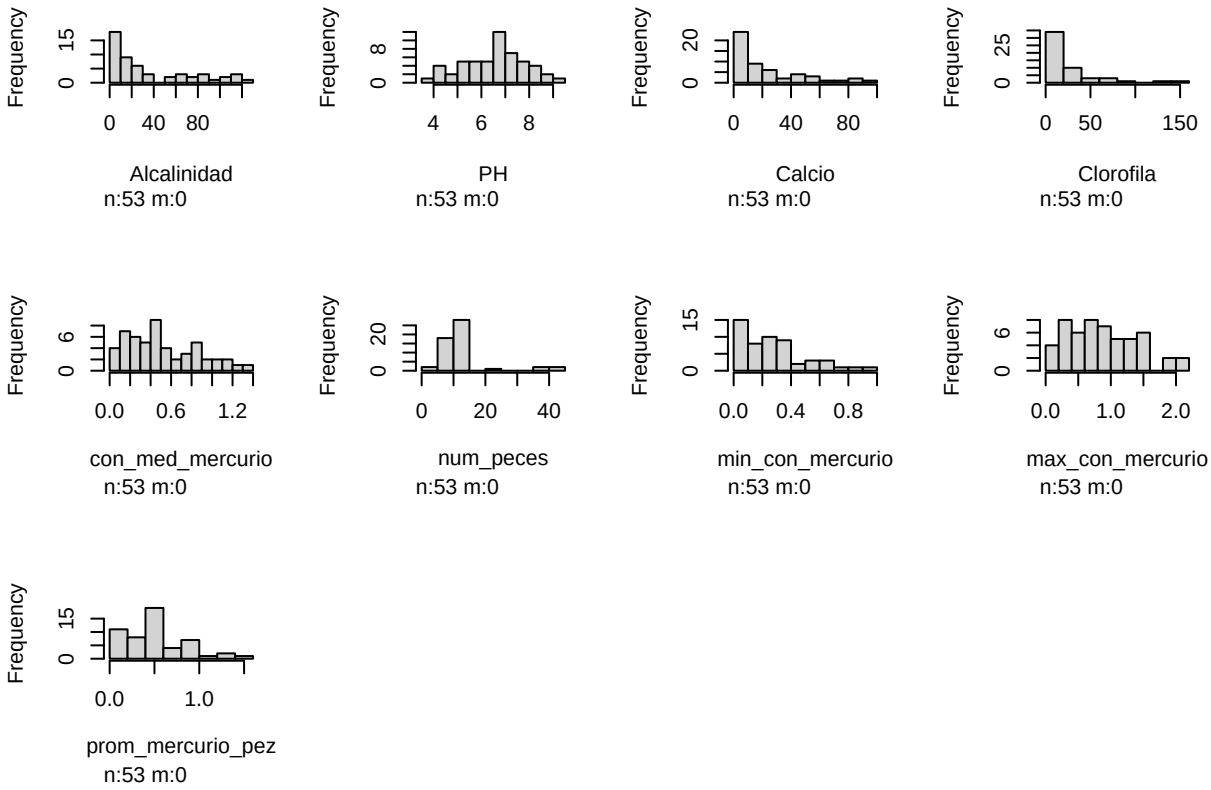



```
## ---------------------------- edad ----------------------------
## Quartil 1:  1   Quartil 3:  1
```



Histogramas

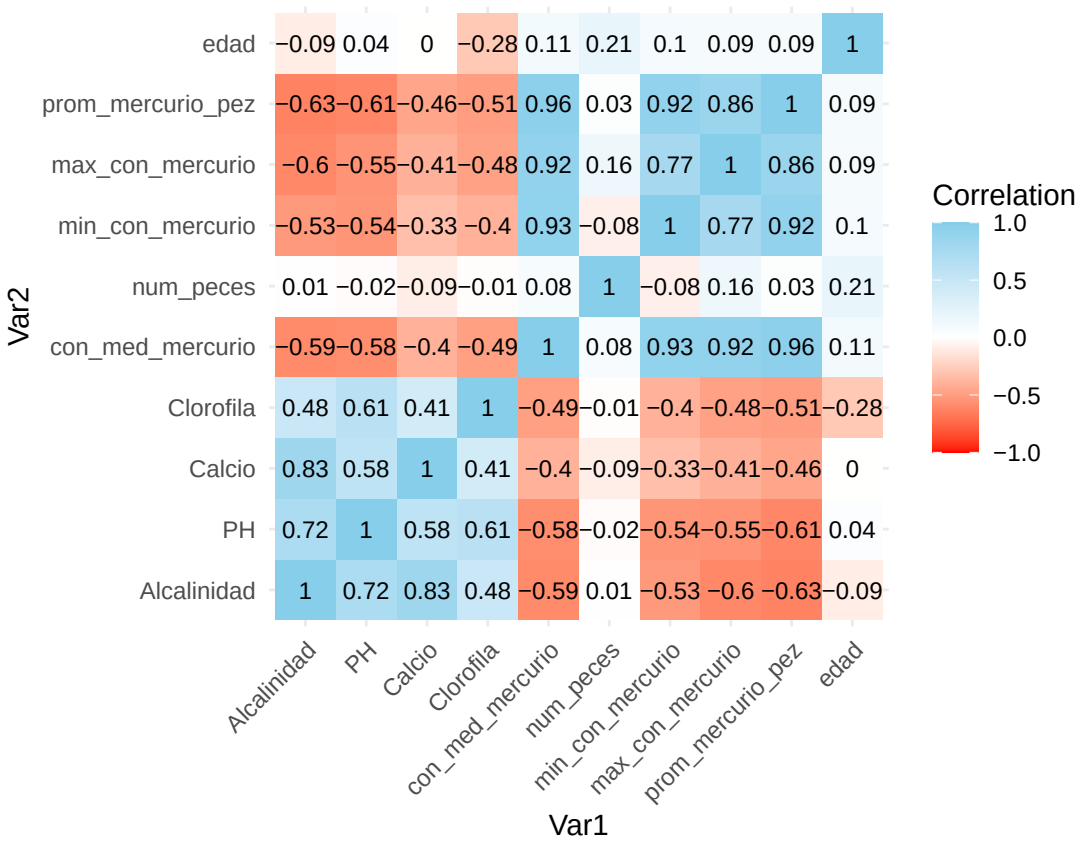
```
data_nums_only <- subset(data, select = -c(Id, Lago));
hist.data.frame(data_nums_only, nclass = 10)
```



Podemos observar que varias de las variables tienen un sesgo a la derecha, lo que indica que la mayoria de los datos se encuentran en la parte izquierda de la distribucion. Esto puede deberse a que los datos fueron tomados de una poblacion que no es normal, o que la muestra no es representativa de la poblacion.

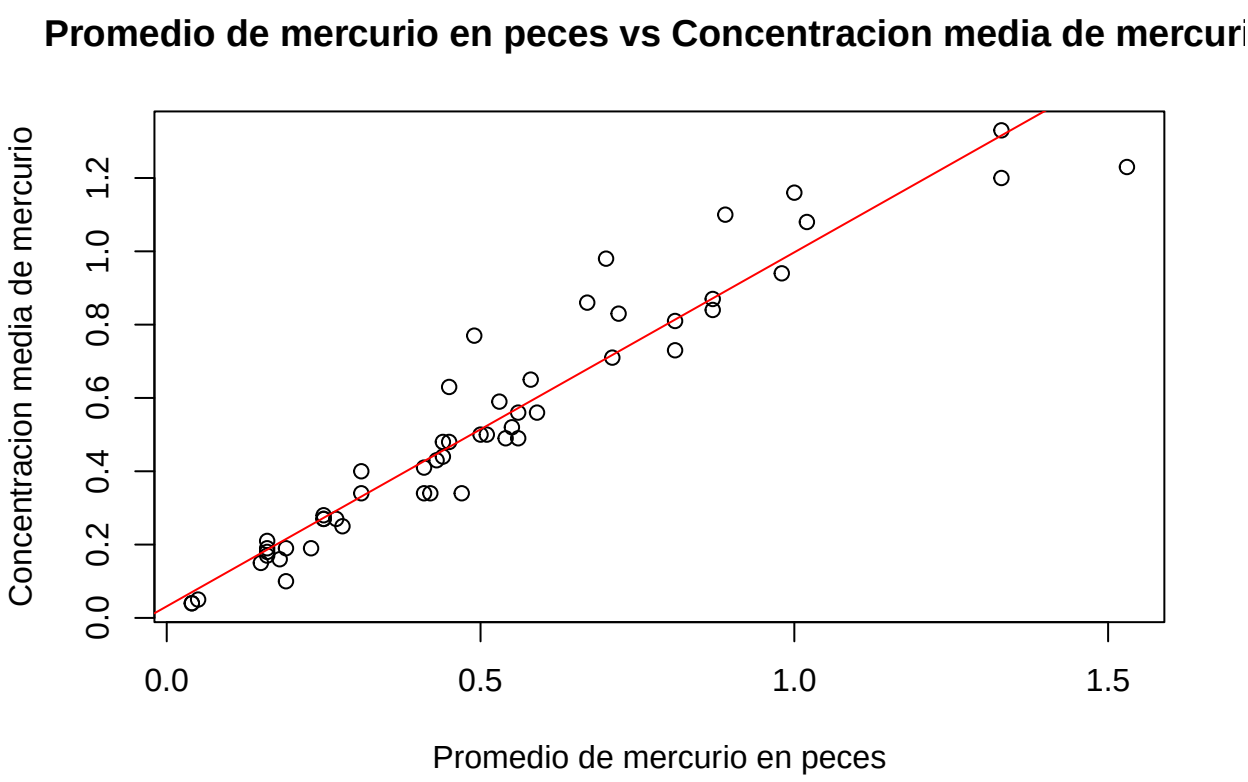
Matriz de correlacion

```
corr_mat <- cor(data_nums_only)
corr_mat <- melt(corr_mat)

ggplot(corr_mat, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  geom_text(aes(label = round(value, 2)), size = 3) +
  scale_fill_gradient2(low = "red", mid = "white", high = "skyblue", midpoint = 0, limit = c(-1, 1), spa
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  coord_fixed()
```

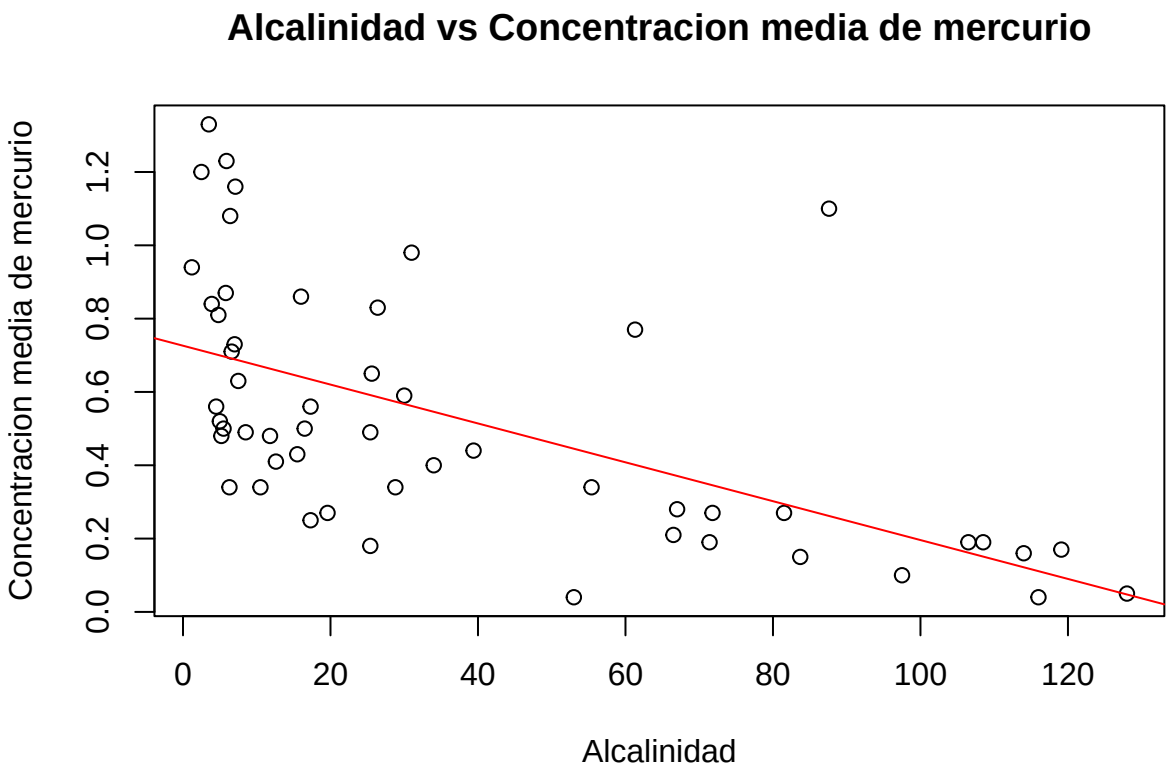


Podemos observar que todas las variables a exepcion de num_peces y edad tienen correlacion moderadas o altas con otras variables. Esto indica que se tendran que eliminar algunas variables para evitar multicolinealidad. Como sabemos que la variable con_med_mercurio es la que queremos predecir, nos quedaremos con las variables que tengan una correlacion alta con esta. Esto nos deja con las variables Alcalinidad, PH, Calcio, Clorofila, min_con_mercurio, max_con_mercurio y prom_mercurio_pez. Analizando las correlaciones entre estas variables, podemos ver que las variables min_con_mercurio, max_con_mercurio y prom_mercurio_pez tienen una correlacion alta entre si, por lo que nos quedaremos con la variable prom_mercurio_pez. Esto nos deja con las variables Alcalinidad, PH, Calcio, Clorofila y prom_mercurio_pez, de las cuales Alcanilidad, PH, Calcio y clorofila tienen una correlacion alta entre si, por lo que nos quedaremos con la variable Alcanilidad. Esto nos deja con las variables Alcalinidad y prom_mercurio_pez. Ya que las variables num_peces y edad tienen una correlacion baja con la variable con_med_mercurio, y no tienen una correlacion alta entre si, las dejaremos para ver si mejoran el modelo.

```
plot(data_nums_only$prom_mercurio_pez, data_nums_only$con_med_mercurio, xlab = "Promedio de mercurio en
abline(lm(data_nums_only$con_med_mercurio ~ data_nums_only$prom_mercurio_pez), col = "red")
```




```
plot(data_nums_only$Alcalinidad, data_nums_only$con_med_mercurio, xlab = "Alcalinidad", ylab = "Concentr
abline(lm(data_nums_only$con_med_mercurio ~ data_nums_only$Alcalinidad), col = "red")
```



Podemos observar que ambas variables tienen una relacion lineal con la variable con_med_mercurio, por lo que las dejaremos en el modelo. Cabe mencionar que Alcanilidad y prom_mercurio_pez tienen una correlacion alta entre si, por lo que se podria eliminar una de las dos variables, pero se dejaran ambas en el modelo para ver si mejora el modelo ya que alcalinidad tiene una relacion negativa con la variable con_med_mercurio, mientras que prom_mercurio_pez tiene una relacion positiva con la variable con_med_mercurio.

Excluimos variables con correlacion alta entre ellas

```
data_clean <- subset(data_nums_only, select=-c(min_con_mercurio, max_con_mercurio, PH, Calcio, Clorofil
```

Regresion lineal multiple con todas las variables seleccionadas

```
rl <- lm(con_med_mercurio ~ ., data = data_clean)
summary(rl)
```

```
##
## Call:
## lm(formula = con_med_mercurio ~ ., data = data_clean)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.266758 -0.039443 -0.004791  0.012307  0.284459
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)       -0.0112483  0.0522008  -0.215    0.830
## Alcalinidad        0.0001120  0.0004596   0.244    0.809
## num_peces          0.0020441  0.0016300   1.254    0.216
## prom_mercurio_pez  0.9711703  0.0517879  18.753   <2e-16 ***
## edad               0.0112349  0.0354797   0.317    0.753
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09832 on 48 degrees of freedom
## Multiple R-squared:  0.9233, Adjusted R-squared:  0.9169
## F-statistic: 144.4 on 4 and 48 DF,  p-value: < 2.2e-16
```

Podemos observar que segun el modelo, la unica variable signficativa es prom_mercurio_pez. Para verificar si esto es cierto utilizaremos a continuacion el metodo de eliminacion hacia ambos lados.

Busqueda del mejor modelo

```
step(rl, direction = "both", trace = 0)
```

```
##
## Call:
## lm(formula = con_med_mercurio ~ prom_mercurio_pez, data = data_clean)
##
## Coefficients:
##       (Intercept)  prom_mercurio_pez
##           0.03154            0.96575
```

Se confirma que la unica variable significativa es prom_mercurio_pez. Esto indica que el modelo con las cuatro variables no es muy bueno, ya que solo una variable explica la variable con_med_mercurio. Por lo que se procedera a eliminar las variables Alcalinidad, edad, y num_peces y se volvera a correr el modelo solo con la variable prom_mercurio_pez.

Regresion lineal con el mejor modelo

```
rl_best <- lm(formula = con_med_mercurio ~ prom_mercurio_pez, data = data_clean)
summary(rl_best)
```

```
##
## Call:
## lm(formula = con_med_mercurio ~ prom_mercurio_pez, data = data_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.27913 -0.04133 -0.01606  0.01402  0.27244
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.03154    0.02444   1.291    0.203
## prom_mercurio_pez  0.96575    0.03985  24.233   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09734 on 51 degrees of freedom
## Multiple R-squared:  0.9201, Adjusted R-squared:  0.9185
## F-statistic: 587.2 on 1 and 51 DF,  p-value: < 2.2e-16
```

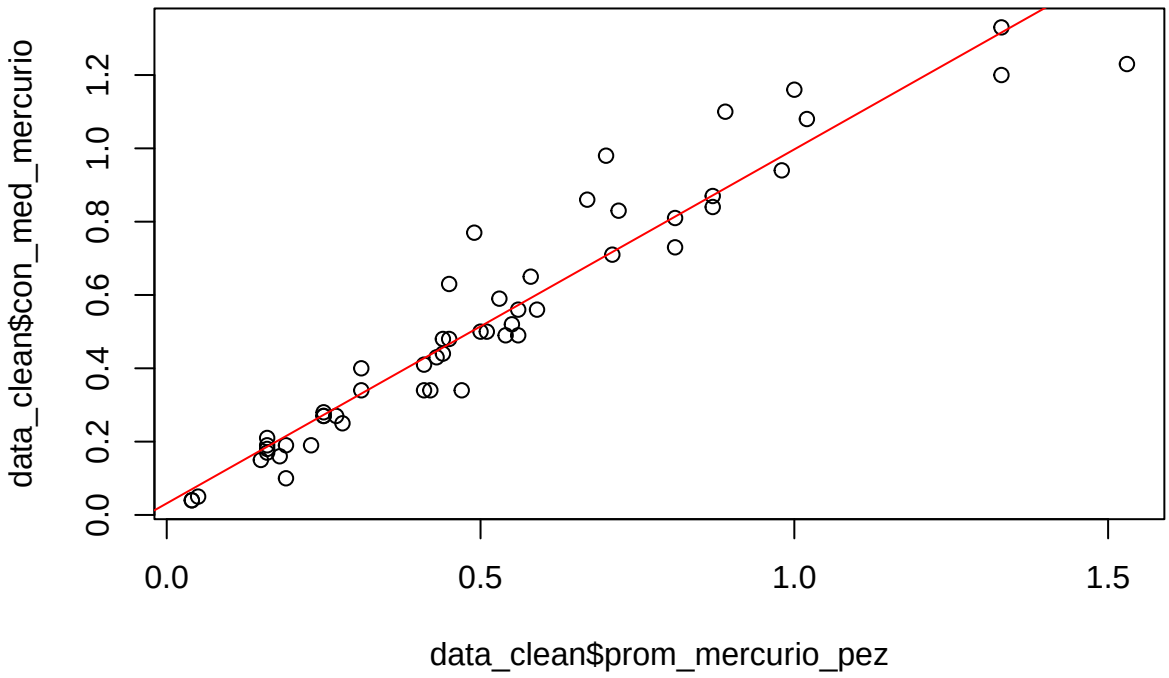
Podemos observar que aunque la R2 bajo ligeramente, la R2 ajustada aumento, lo que indica que el modelo es mejor que el anterior. Tambien podemos observar que la variable prom_mercurio_pez se mantiene como significativa.

Ecuacion de la regresion lineal

```
cat("con_med_mercurio = ", round(rl_best$coefficients[1], 4), " +",
    round(rl_best$coefficients[2], 4), "* prom_mercurio_pez")
```

```
## con_med_mercurio =  0.0315  + 0.9657 * prom_mercurio_pez
```

```
plot(data_clean$prom_mercurio_pez, data_clean$con_med_mercurio)
abline(rl_best, col = "red")
```



Validacion del modelo

Pruebas de hipotesis

Hipotesis: h0: beta1 = 0 h1: beta1 != 0

Reglas de decision:
* Si p-value < alpha, se rechaza H0 y se acepta H1
* Si p-value > alpha, se rechaza H1 y se acepta H0
* Si t* > t, se rechaza H0 y se acepta H1
* Si t* < t, se rechaza H1 y se acepta H0

```r
s <- summary(rl_best)
alpha <- 0.05
n <- nrow(data_nums_only)
t0 <- abs(qt(alpha / 2, n - 2))
tes <- s$coefficients[, 3]
for (i in 2:(length(tes))) {
  if (tes[i] > t0 & s$coefficients[i, 4] < alpha) {
    cat("La variable", names(rl_best$coefficients)[i], "es significativa. (t* > t0 & p < alpha)\n",
        "t* =", round(tes[i], 4), ", t0 =", round(t0, 4), "\n",
        "p-value =", s$coefficients[i, 4], ", alpha =", alpha, "\n")
  } else {
    cat("La variable", names(rl_best$coefficients)[i], "no es significativa. (t* < t0 & p > alpha)\n",
        "t* =", round(tes[i], 4), ", t0 =", round(t0, 4), "\n",
        "p-value =", s$coefficients[i, 4], ", alpha =", alpha, "\n")
  }
}
```

```
## La variable prom_mercurio_pez es significativa. (t* > t0 & p < alpha)
##  t* = 24.2331 , t0 = 2.0076
##  p-value = 1.200347e-29 , alpha = 0.05
```

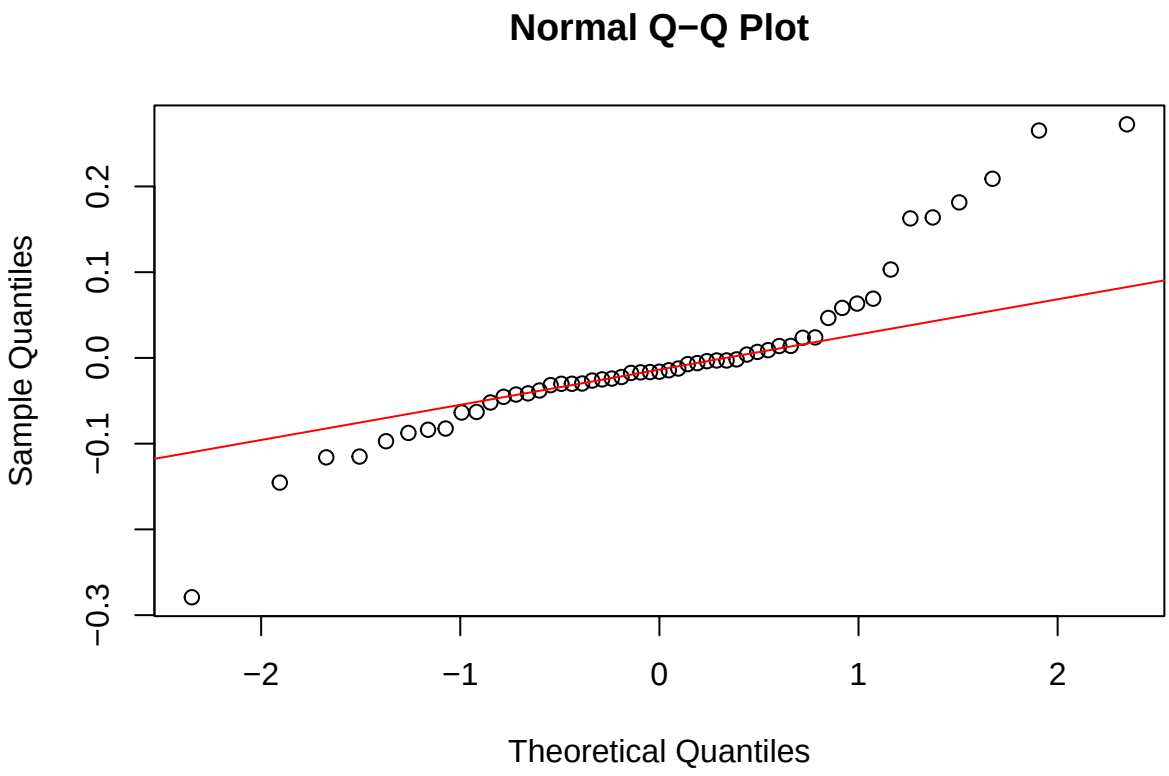
En este caso al solo tener una variable independiente, solo existe la hipotesis para B1. Como podemos observar, la variable prom_mercurio_pez es significativa, ya que el p-value es menor que alpha y la t* es mayor que t0. Confirmando asi que la variable prom_mercurio_pez es significativa para explicar la variable con_med_mercurio.

Verificación de supuestos

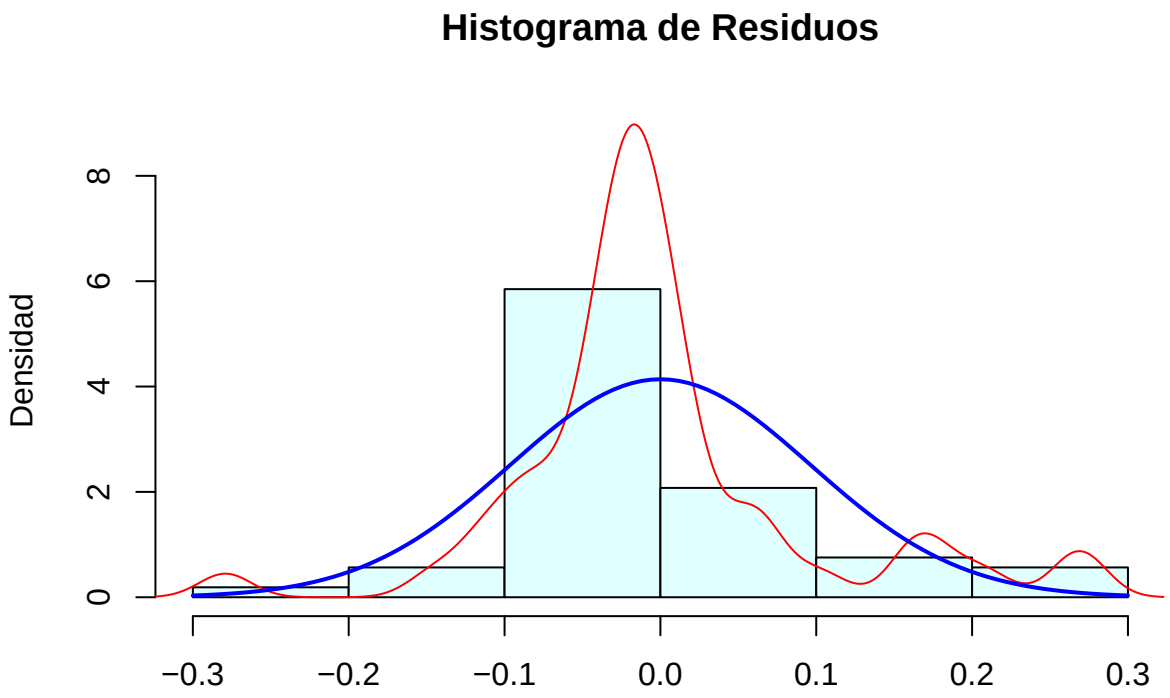
Normalidad de los residuos

Hipotesis:
* H0: miu = 0
* H1: miu != 0
Reglas de decision:
* Si p-value < alpha, se rechaza H0 y se acepta H1
* Si p-value > alpha, se rechaza H1 y se acepta H0

```r
E<-rl_best$residuals
Y<-rl_best$fitted.values

qqnorm(E)
qqline(E,col="red")
```




```
hist(E,col="lightcyan",freq=FALSE,main="Histograma de Residuos",xlab="",ylab="Densidad", ylim=c(0, max(
lines(density(E),col="red")
curve(dnorm(x,mean=mean(E),sd=sd(E)), add=TRUE, col="blue",lwd=2)
```




```
shapiro.test(E)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  E
## W = 0.89095, p-value = 0.0001611
```

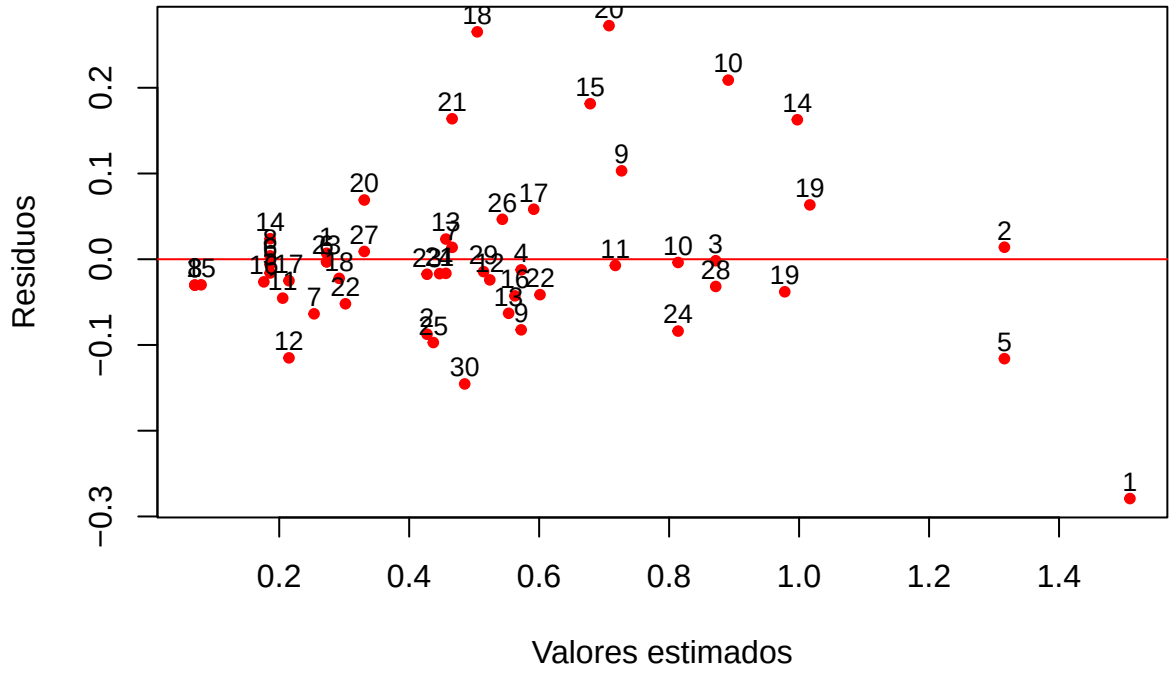
```
t.test(E, alternative = "two.sided")
```

```
##
##  One Sample t-test
##
## data:  E
## t = -1.0135e-15, df = 52, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.02657205  0.02657205
## sample estimates:
##     mean of x
## -1.342143e-17
```

Podemos observar en el qqplot que los residuos siguen una distribución con colas gruesas. Tambien podemos observar que el p-value de la prueba de shapiro es menor que alpha, por lo que podemos rechazar la hipotesis nula y decir que los residuos no siguen una distribución normal. Finalmente podemos observar que la prueba t nos da una media diferente de cero, lo que tambien no nos permite rechazar la hipotesis nula.

Homocedasticidad y modelo apropiado

```
plot(Y,E,ylab="Residuos",xlab="Valores estimados",pch=20,col="red")
abline(h=0,col="red")
text(Y[],E[],1:30,cex=0.8,pos=3,offset=0.2)
```

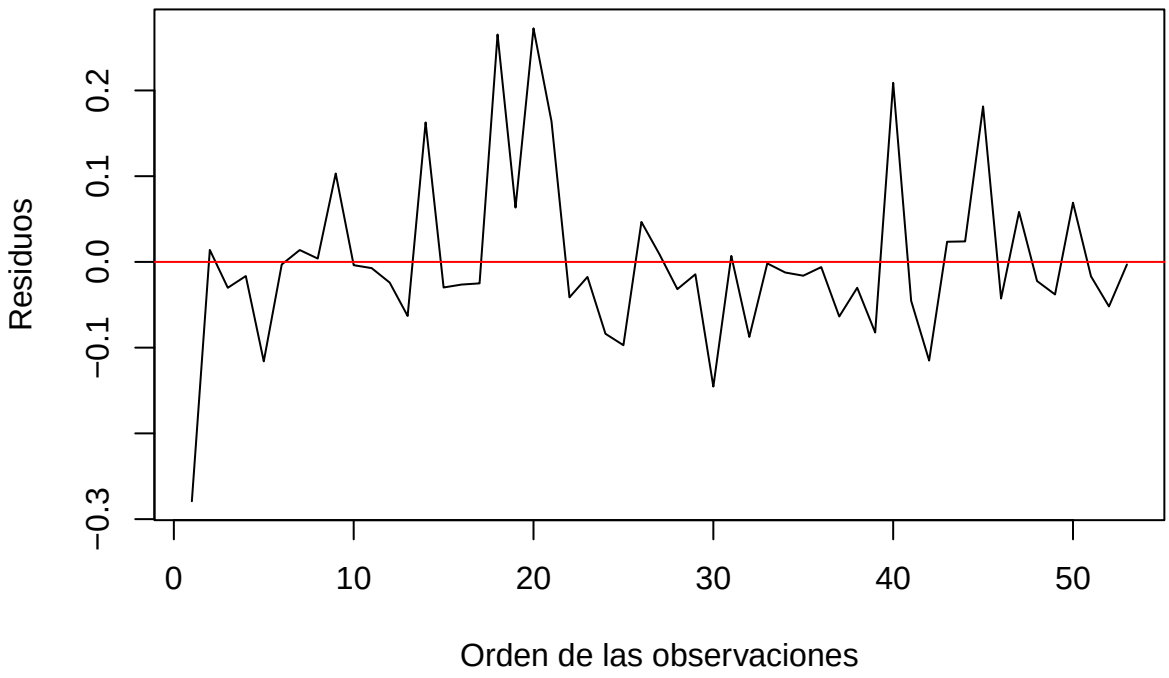


En la grafica podemos observar que los residuos no aparentan seguir algun tipo de patron, por lo que podemos decir que los residuos son homocedasticos y que el modelo es apropiado.

Independencia

Hipotesis:
* H0: rho = 0
* H1: rho != 0
Reglas de decision:
* Si p-value < alpha, se rechaza H0 y se acepta H1
* Si p-value > alpha, se rechaza H1 y se acepta H0

```
n<-length(data_clean$con_med_mercurio)
plot(c(1:n),rl_best$residuals,type="l",xlab="Orden de las observaciones",ylab="Residuos")
abline(h=0,col="red")
```




```
dwt(rl_best,alternative="two.sided")
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.06460033      1.709555   0.272
##  Alternative hypothesis: rho != 0
```

Podemos observar que los residuos no siguen un patron, por lo que podemos decir que los residuos son independientes. Tambien podemos observar que el p-value de la prueba de durbin watson es mayor que alpha, por lo que podemos aceptar la hipotesis nula y decir que los residuos son independientes.

Conclusiones

¿Cuáles son los principales factores que influyen en el nivel de contaminación por mercurio en los peces de los lagos de Florida?
Tras realizar el analisis de regresion lineal, podemos decir que el principal factor que influye en el nivel de contaminacion por mercurio en los peces de los lagos de Florida es el promedio de mercurio en los peces de los lagos.
Ademas de esto podemos concluir que tanto el promedio como el maximo de mercurio en los peces de los lagos serian signficativos dependiendo de lo que se quiera analizar.
Esto se debe a que ambas variables tienen una correlacion alta con la variable dependiente y entre si, posiblemente resultando en modelos de regresion lineal similares al utilizarse individualmente.