

# Técnicas de procesamiento de datos para el análisis estadístico y para la construcción de modelos

Emilia Victoria Jácome Iñiguez

2022-09-05

La contaminación por mercurio de peces en el agua dulce comestibles es una amenaza directa contra nuestra salud. Se llevó a cabo un estudio reciente en 53 lagos de Florida con el fin de examinar los factores que influían en el nivel de contaminación por mercurio. Las variables que se midieron se encuentran en mercurio.csv

Alrededor de la principal pregunta de investigación que surge en este estudio: # 0. Pregunta Base: ¿Cuáles son los principales factores que influyen en el nivel de contaminación por mercurio en los peces de los lagos de Florida? pueden surgir preguntas paralelas que desglosan esta pregunta general:

¿Hay evidencia para suponer que la concentración promedio de mercurio en los lagos es dañino para la salud humana?

Considera que las normativas de referencia para evaluar los niveles máximos de Hg (Reglamento 34687-MAG y los reglamentos internacionales CE 1881/2006 y Codex Standard 193-1995) establecen que la concentración promedio de mercurio en productos de la pesca no debe superar los 0.5 mg de Hg/kg.

¿Habrá diferencia significativa entre la concentración de mercurio por la edad de los peces?

Si el muestreo se realizó lanzando una red y analizando los peces que la red encontraba ¿Habrá influencia del número de peces encontrados en la concentración de mercurio en los peces?

¿Las concentraciones de alcalinidad, clorofila, calcio en el agua del lago influyen en la concentración de mercurio de los peces?

## 1. EXPLORACIÓN DE LA BASE DE DATOS

Accede a la base de datos de las mediciones del estudio de Mercurio en los lagos de Florida: Aquí Descargar Aquí.

```
df=read.csv("mercurio.csv") #Leer la base de datos
df
```

##	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12
## 1	1	Alligator	5.9	6.1	3.0	0.7	1.23	5	0.85	1.43	1.53	1
## 2	2	Annie	3.5	5.1	1.9	3.2	1.33	7	0.92	1.90	1.33	0
## 3	3	Apopka	116.0	9.1	44.1	128.3	0.04	6	0.04	0.06	0.04	0
## 4	4	Blue Cypress	39.4	6.9	16.4	3.5	0.44	12	0.13	0.84	0.44	0
## 5	5	Brick	2.5	4.6	2.9	1.8	1.20	12	0.69	1.50	1.33	1
## 6	6	Bryant	19.6	7.3	4.5	44.1	0.27	14	0.04	0.48	0.25	1
## 7	7	Cherry	5.2	5.4	2.8	3.4	0.48	10	0.30	0.72	0.45	1
## 8	8	Crescent	71.4	8.1	55.2	33.7	0.19	12	0.08	0.38	0.16	1
## 9	9	Deer Point	26.4	5.8	9.2	1.6	0.83	24	0.26	1.40	0.72	1
## 10	10	Dias	4.8	6.4	4.6	22.5	0.81	12	0.41	1.47	0.81	1
## 11	11	Dorr	6.6	5.4	2.7	14.9	0.71	12	0.52	0.86	0.71	1
## 12	12	Down	16.5	7.2	13.8	4.0	0.50	12	0.10	0.73	0.51	1
## 13	13	Eaton	25.4	7.2	25.2	11.6	0.49	7	0.26	1.01	0.54	1
## 14	14	East Tohopekaliga	7.1	5.8	5.2	5.8	1.16	43	0.50	2.03	1.00	1
## 15	15	Farm-13	128.0	7.6	86.5	71.1	0.05	11	0.04	0.11	0.05	0
## 16	16	George	83.7	8.2	66.5	78.6	0.15	10	0.12	0.18	0.15	1
## 17	17	Griffin	108.5	8.7	35.6	80.1	0.19	40	0.07	0.43	0.19	1
## 18	18	Harney	61.3	7.8	57.4	13.9	0.77	6	0.32	1.50	0.49	1
## 19	19	Hart	6.4	5.8	4.0	4.6	1.08	10	0.64	1.33	1.02	1
## 20	20	Hatchineha	31.0	6.7	15.0	17.0	0.98	6	0.67	1.44	0.70	1
## 21	21	Iamonia	7.5	4.4	2.0	9.6	0.63	12	0.33	0.93	0.45	1
## 22	22	Istokpoga	17.3	6.7	10.7	9.5	0.56	12	0.37	0.94	0.59	1
## 23	23	Jackson	12.6	6.1	3.7	21.0	0.41	12	0.25	0.61	0.41	0
## 24	24	Josephine	7.0	6.9	6.3	32.1	0.73	12	0.33	2.04	0.81	1
## 25	25	Kingsley	10.5	5.5	6.3	1.6	0.34	10	0.25	0.62	0.42	1
## 26	26	Kissimmee	30.0	6.9	13.9	21.5	0.59	36	0.23	1.12	0.53	1
## 27	27	Lochloosa	55.4	7.3	15.9	24.7	0.34	10	0.17	0.52	0.31	1
## 28	28	Louisa	3.9	4.5	3.3	7.0	0.84	8	0.59	1.38	0.87	1
## 29	29	Miccasukee	5.5	4.8	1.7	14.8	0.50	11	0.31	0.84	0.50	0
## 30	30	Minneola	6.3	5.8	3.3	0.7	0.34	10	0.19	0.69	0.47	1
## 31	31	Monroe	67.0	7.8	58.6	43.8	0.28	10	0.16	0.59	0.25	1
## 32	32	Newmans	28.8	7.4	10.2	32.7	0.34	10	0.16	0.65	0.41	1
## 33	33	Ocean Pond	5.8	3.6	1.6	3.2	0.87	12	0.31	1.90	0.87	0
## 34	34	Ocheese Pond	4.5	4.4	1.1	3.2	0.56	13	0.25	1.02	0.56	0
## 35	35	Okeechobee	119.1	7.9	38.4	16.1	0.17	12	0.07	0.30	0.16	1
## 36	36	Orange	25.4	7.1	8.8	45.2	0.18	13	0.09	0.29	0.16	1
## 37	37	Panasoffkee	106.5	6.8	90.7	16.5	0.19	13	0.05	0.37	0.23	1
## 38	38	Parker	53.0	8.4	45.6	152.4	0.04	4	0.04	0.06	0.04	0
## 39	39	Placid	8.5	7.0	2.5	12.8	0.49	12	0.31	0.63	0.56	1
## 40	40	Puzzle	87.6	7.5	85.5	20.1	1.10	10	0.79	1.41	0.89	1
## 41	41	Rodman	114.0	7.0	72.6	6.4	0.16	14	0.04	0.26	0.18	1
## 42	42	Rousseau	97.5	6.8	45.5	6.2	0.10	12	0.05	0.26	0.19	1
## 43	43	Sampson	11.8	5.9	24.2	1.6	0.48	10	0.27	1.05	0.44	1
## 44	44	Shipp	66.5	8.3	26.0	68.2	0.21	12	0.05	0.48	0.16	1
## 45	45	Talquin	16.0	6.7	41.2	24.1	0.86	12	0.36	1.40	0.67	1
## 46	46	Tarpon	5.0	6.2	23.6	9.6	0.52	12	0.31	0.95	0.55	1
## 47	51	Tohopekaliga	25.6	6.2	12.6	27.7	0.65	44	0.30	1.10	0.58	1
## 48	47	Trafford	81.5	8.9	20.5	9.6	0.27	6	0.04	0.40	0.27	0
## 49	48	Trout	1.2	4.3	2.1	6.4	0.94	10	0.59	1.24	0.98	1
## 50	49	Tsala Apopka	34.0	7.0	13.1	4.6	0.40	12	0.08	0.90	0.31	1
## 51	50	Weir	15.5	6.9	5.2	16.5	0.43	11	0.23	0.69	0.43	1
## 52	52	Wildcat	17.3	5.2	3.0	2.6	0.25	12	0.15	0.40	0.28	1
## 53	53	Yale	71.8	7.9	20.5	8.8	0.27	12	0.15	0.51	0.25	1

##Explora las variables y familiarízate con su significado.

La descripción de cada variable es la siguiente:

X1 = número de indentificación X2 = nombre del lago X3 = alcalinidad (mg/l de carbonato de calcio) X4 = PH X5 = calcio (mg/l) X6 = clorofila (mg/l) X7 = concentración media de mercurio (parte por millón) en el tejido muscular del grupo de peces estudiados en cada lago X8 = número de peces estudiados en el lago X9 = mínimo de la concentración de mercurio en cada grupo de peces X10 = máximo de la concentración de mercurio en cada grupo de peces X11 = estimación (mediante regresión) de la concentración de mercurio en el pez de 3 años (o promedio de mercurio cuando la edad no está disponible) X12 = indicador de la edad de los peces (0: jóvenes; 1: maduros)

###Identifica la cantidad de datos y variables presentes.

```
n_col =length(df)
print(paste("La cantidad de columnas es: ", n_col))
```

```
## [1] "La cantidad de columnas es:  12"
```

```
n = nrow(df)
print(paste("La cantidad de datos es de: ", n))
```

```
## [1] "La cantidad de datos es de:  53"
```

###Clasifica las variables de acuerdo a su tipo y escala de medición.

```
str(df)
```

```
## 'data.frame':   53 obs. of  12 variables:
## $ X1 : int  1 2 3 4 5 6 7 8 9 10 ...
## $ X2 : chr  "Alligator" "Annie" "Apopka" "Blue Cypress" ...
## $ X3 : num  5.9 3.5 116 39.4 2.5 19.6 5.2 71.4 26.4 4.8 ...
## $ X4 : num  6.1 5.1 9.1 6.9 4.6 7.3 5.4 8.1 5.8 6.4 ...
## $ X5 : num  3 1.9 44.1 16.4 2.9 4.5 2.8 55.2 9.2 4.6 ...
## $ X6 : num  0.7 3.2 128.3 3.5 1.8 ...
## $ X7 : num  1.23 1.33 0.04 0.44 1.2 0.27 0.48 0.19 0.83 0.81 ...
## $ X8 : int  5 7 6 12 12 14 10 12 24 12 ...
## $ X9 : num  0.85 0.92 0.04 0.13 0.69 0.04 0.3 0.08 0.26 0.41 ...
## $ X10: num  1.43 1.9 0.06 0.84 1.5 0.48 0.72 0.38 1.4 1.47 ...
## $ X11: num  1.53 1.33 0.04 0.44 1.33 0.25 0.45 0.16 0.72 0.81 ...
## $ X12: int  1 0 0 0 1 1 1 1 1 1 ...
```

```
#Set de variables numéricas
df_num = df[,c(-1,-2, -12)]
```

```
#Set de variable categóricas
df_cat = df[,c(1,2,12)]
```

##Exploración de la base de datos - Medidas estadísticas ### Variables cuantitativas #####Medidas de tendencia Central

```
summary(df)
```

##	X1	X2	X3	X4
##	Min. : 1	Length:53	Min. : 1.20	Min. :3.600
##	1st Qu.:14	Class :character	1st Qu.: 6.60	1st Qu.:5.800
##	Median :27	Mode :character	Median : 19.60	Median :6.800
##	Mean :27		Mean : 37.53	Mean :6.591
##	3rd Qu.:40		3rd Qu.: 66.50	3rd Qu.:7.400
##	Max. :53		Max. :128.00	Max. :9.100
##	X5	X6	X7	X8
##	Min. : 1.1	Min. : 0.70	Min. :0.0400	Min. : 4.00
##	1st Qu.: 3.3	1st Qu.: 4.60	1st Qu.:0.2700	1st Qu.:10.00
##	Median :12.6	Median : 12.80	Median :0.4800	Median :12.00
##	Mean :22.2	Mean : 23.12	Mean :0.5272	Mean :13.06
##	3rd Qu.:35.6	3rd Qu.: 24.70	3rd Qu.:0.7700	3rd Qu.:12.00
##	Max. :90.7	Max. :152.40	Max. :1.3300	Max. :44.00
##	X9	X10	X11	X12
##	Min. :0.0400	Min. :0.0600	Min. :0.0400	Min. :0.0000
##	1st Qu.:0.0900	1st Qu.:0.4800	1st Qu.:0.2500	1st Qu.:1.0000
##	Median :0.2500	Median :0.8400	Median :0.4500	Median :1.0000
##	Mean :0.2798	Mean :0.8745	Mean :0.5132	Mean :0.8113
##	3rd Qu.:0.3300	3rd Qu.:1.3300	3rd Qu.:0.7000	3rd Qu.:1.0000
##	Max. :0.9200	Max. :2.0400	Max. :1.5300	Max. :1.0000

Cálculo de moda para cada variable

```
#install.packages("modeest")
library(modeest)

## Warning: package 'modeest' was built under R version 4.1.3

apply(df[,c(-1,-2)], 2, mlv, method = "mfv")
```

```
## $X3
## [1] 17.3 25.4
##
## $X4
## [1] 5.8 6.9
##
## $X5
## [1] 3.0 3.3 5.2 6.3 20.5
##
## $X6
## [1] 1.6 3.2 9.6
##
## $X7
## [1] 0.34
##
## $X8
## [1] 12
##
## $X9
## [1] 0.04
##
## $X10
## [1] 0.06 0.26 0.40 0.48 0.69 0.84 1.40 1.50 1.90
##
## $X11
## [1] 0.16
##
## $X12
## [1] 1
```

*#Se quitaron las primeras variables ya que no aportan información para la moda, ya que cada una es única.*

#### ####Medidas de Dispersión

```
print(paste("Desviación estándar X3: ", sd(df$X3)))
```

```
## [1] "Desviación estándar X3: 38.2035267446992"
```

```
print(paste("Varianza X3: ", var(df$X3)))
```

```
## [1] "Varianza X3: 1459.50945573295"
```

```
print(paste("Rango X3: ", diff(range(df$X3))))
```

```
## [1] "Rango X3: 126.8"
```

```
print(paste("Desviación estándar X4: ", sd(df$X4)))
```

```
## [1] "Desviación estándar X4: 1.28844929916419"
```

```
print(paste("Varianza X4: ", var(df$X4)))
```

```
## [1] "Varianza X4:  1.66010159651669"
```

```
print(paste("Rango X4: ", diff(range(df$X4))))
```

```
## [1] "Rango X4:  5.5"
```

```
print(paste("Desviación estándar X5: ", sd(df$X5)))
```

```
## [1] "Desviación estándar X5:  24.9325743877828"
```

```
print(paste("Varianza X5: ", var(df$X5)))
```

```
## [1] "Varianza X5:  621.633265602322"
```

```
print(paste("Rango X5: ", diff(range(df$X5))))
```

```
## [1] "Rango X5:  89.6"
```

```
print(paste("Desviación estándar X6: ", sd(df$X6)))
```

```
## [1] "Desviación estándar X6:  30.8163214487754"
```

```
print(paste("Varianza X6: ", var(df$X6)))
```

```
## [1] "Varianza X6:  949.645667634253"
```

```
print(paste("Rango X6: ", diff(range(df$X6))))
```

```
## [1] "Rango X6:  151.7"
```

```
print(paste("Desviación estándar X7: ", sd(df$X7)))
```

```
## [1] "Desviación estándar X7:  0.341035625019815"
```

```
print(paste("Varianza X7: ", var(df$X7)))
```

```
## [1] "Varianza X7:  0.116305297532656"
```

```
print(paste("Rango X7: ", diff(range(df$X7))))
```

```
## [1] "Rango X7:  1.29"
```

```
print(paste("Desviación estándar X8: ", sd(df$X8)))
```

```
## [1] "Desviación estándar X8:  8.56067730592267"
```

```
print(paste("Varianza X8: ", var(df$X8)))
```

```
## [1] "Varianza X8:  73.2851959361393"
```

```
print(paste("Rango X8: ", diff(range(df$X8))))
```

```
## [1] "Rango X8:  40"
```

```
print(paste("Desviación estándar X9: ", sd(df$X9)))
```

```
## [1] "Desviación estándar X9:  0.226405784157881"
```

```
print(paste("Varianza X9: ", var(df$X9)))
```

```
## [1] "Varianza X9:  0.0512595791001451"
```

```
print(paste("Rango X9: ", diff(range(df$X9))))
```

```
## [1] "Rango X9:  0.88"
```

```
print(paste("Desviación estándar X10: ", sd(df$X10)))
```

```
## [1] "Desviación estándar X10:  0.522046881322917"
```

```
print(paste("Varianza X10: ", var(df$X10)))
```

```
## [1] "Varianza X10:  0.272532946298984"
```

```
print(paste("Rango X10: ", diff(range(df$X10))))
```

```
## [1] "Rango X10:  1.98"
```

```
print(paste("Desviación estándar X11: ", sd(df$X11)))
```

```
## [1] "Desviación estándar X11:  0.338729376805697"
```

```
print(paste("Varianza X11: ", var(df$X11)))
```

```
## [1] "Varianza X11: 0.114737590711176"
```

```
print(paste("Rango X11: ", diff(range(df$X11))))
```

```
## [1] "Rango X11: 1.49"
```

```
print(paste("Desviación estándar X12: ", sd(df$X12)))
```

```
## [1] "Desviación estándar X12: 0.394997749437839"
```

```
print(paste("Varianza X12: ", var(df$X12)))
```

```
## [1] "Varianza X12: 0.156023222060958"
```

```
print(paste("Rango X12: ", diff(range(df$X12))))
```

```
## [1] "Rango X12: 1"
```

###Variables cualitativas ###Tabla de distribución de frecuencia

```
print("Tabla de frecuencias para la variable X2 (Nombre de Lago)")
```

```
## [1] "Tabla de frecuencias para la variable X2 (Nombre de Lago)"
```

```
table(df$X2)
```



```
##
##      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
```

```
print("Tabla de frecuencias para la variable X12 (Edad de clasificación de los peses 0: Jóvenes, 1: Adultos)")
```

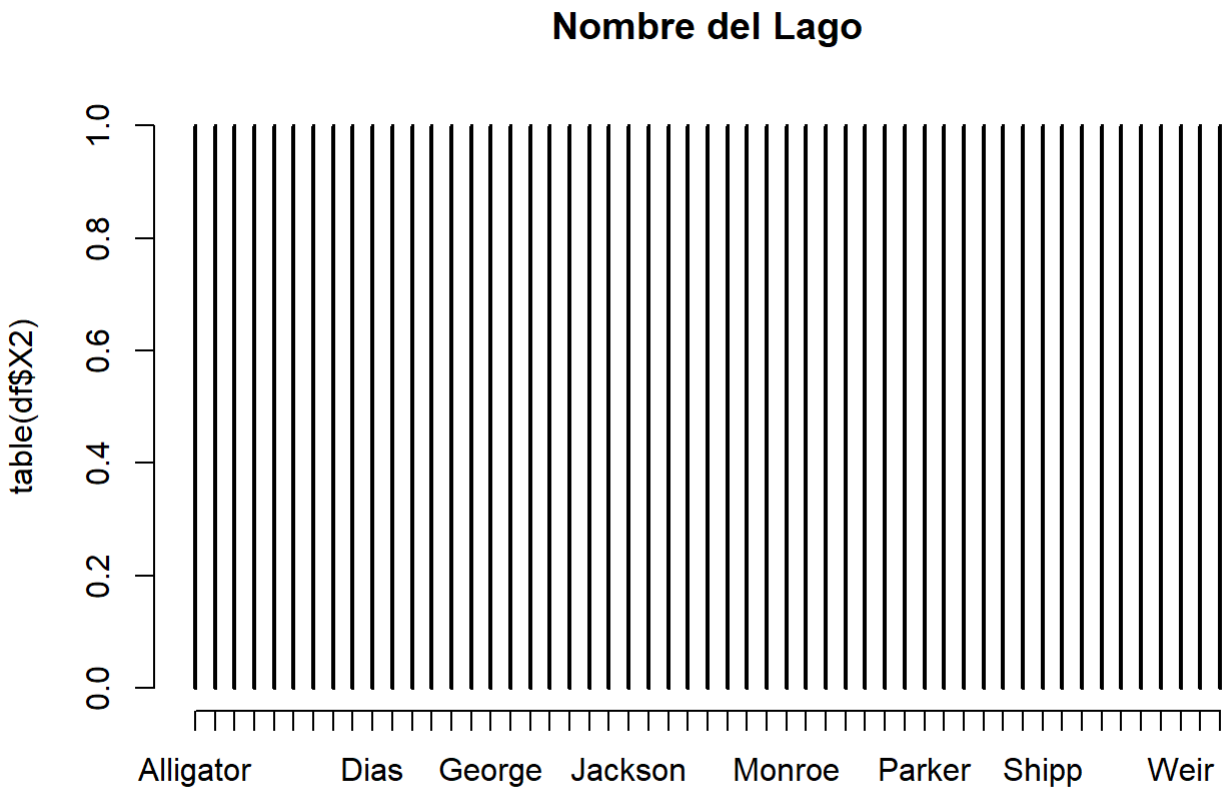
```
## [1] "Tabla de frecuencias para la variable X12 (Edad de clasificación de los peses 0: Jóvenes, 1: Adultos)"
```

```
table(df$X12)
```

```
##
##  0  1
## 10 43
```

###Moda

```
plot(table(df$X2), main ="Nombre del Lago")
```



```
plot(table(df$X12), main ="Edad de los peces")
```

Edad de los peces



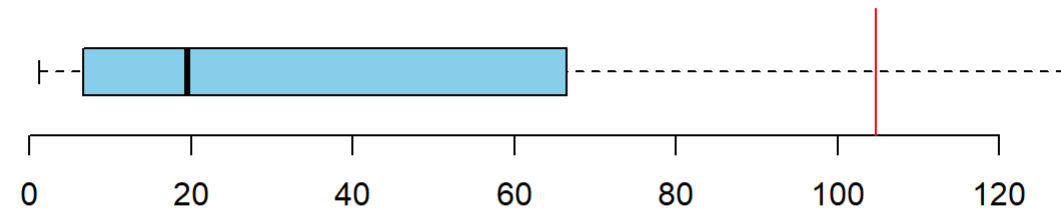
Se puede ver que para la variable X2

(Nombre de Lago) no existe un valor de moda ya que todos los valores son únicos y diferentes haciendo referencia a cada lago en estudio, mientras que para la variable X12, está muy claro por la gráfica que la moda que predomina en la mayoría de lagos, son los peces adultos.

##Explora los datos usando herramientas de visualización ###Variables cuantitativas:

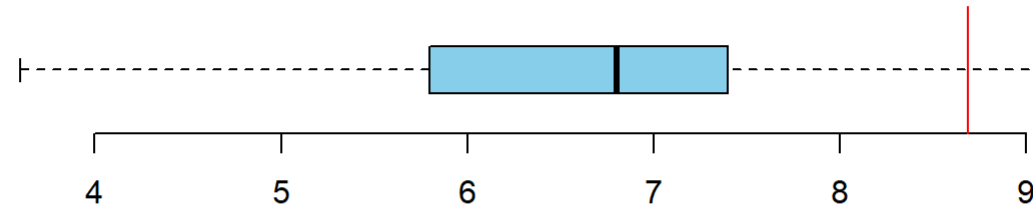
```
col = df$X3
# Para columna X3
cuartiles_X3 = quantile(col, c(0.25, 0.5, 0.75), type = 6);

par(mfrow=c(2,1)) #Matriz de gráficos de 12x2
boxplot(col, col="skyblue", horizontal = TRUE, frame.plot=F, main = "X3")
value = quantile(col,0.75)+sd(col)
abline(v=value,col="red") #Linea vertical en el límite de Los datos atípicos
```

**X3**

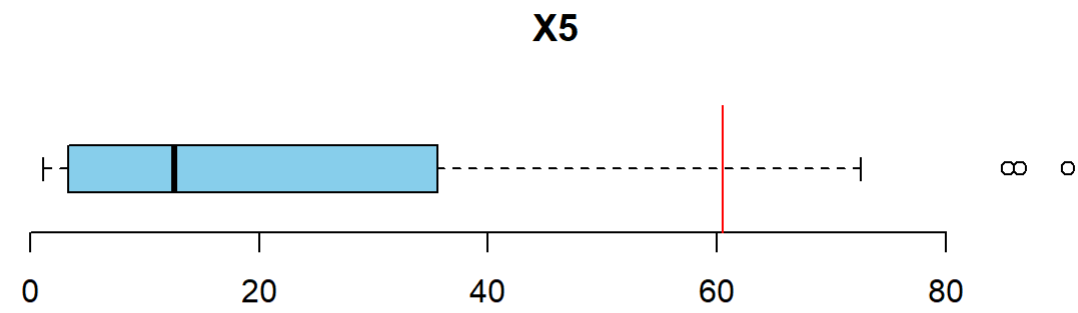
```
col = df$X4
# Para columna X4
cuartiles_X4 = quantile(col, c(0.25, 0.5, 0.75), type = 6);

par(mfrow=c(2,1)) #Matriz de gráficos de 12x2
boxplot(col, col="skyblue", horizontal = TRUE, frame.plot=F, main = "X4")
value = quantile(col,0.75)+sd(col)
abline(v=value,col="red") #Línea vertical en el límite de los datos atípicos
```

**X4**

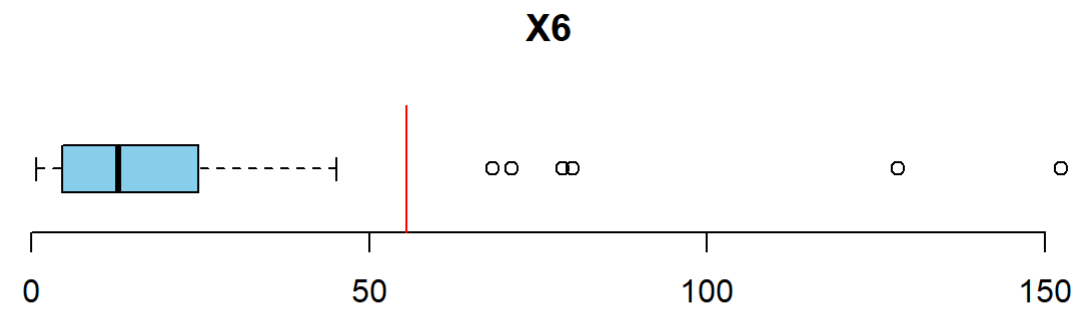
```
col = df$X5
# Para columna X5
cuartiles_X5 = quantile(col, c(0.25, 0.5, 0.75), type = 6);

par(mfrow=c(2,1)) #Matriz de gráficos de 12x2
boxplot(col, col="skyblue", horizontal = TRUE, frame.plot=F, main = "X5")
value = quantile(col,0.75)+sd(col)
abline(v=value,col="red") #Línea vertical en el límite de los datos atípicos
```



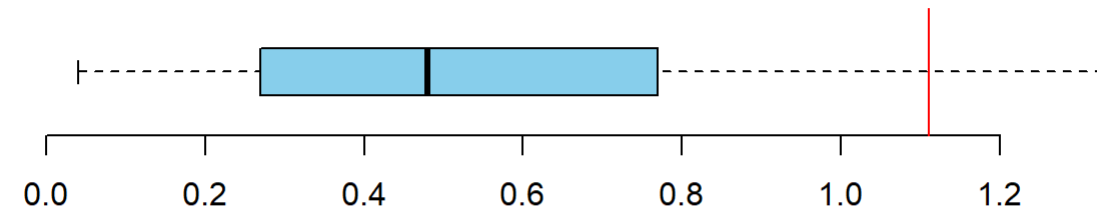
```
col = df$X6
# Para columna X6
cuartiles_X6 = quantile(col, c(0.25, 0.5, 0.75), type = 6);

par(mfrow=c(2,1)) #Matriz de gráficos de 12x2
boxplot(col, col="skyblue", horizontal = TRUE, frame.plot=F, main = "X6")
value = quantile(col,0.75)+sd(col)
abline(v=value,col="red") #Linea vertical en el límite de los datos atípicos
```



```
col = df$X7
# Para columna X7
cuartiles_X7 = quantile(col, c(0.25, 0.5, 0.75), type = 6);

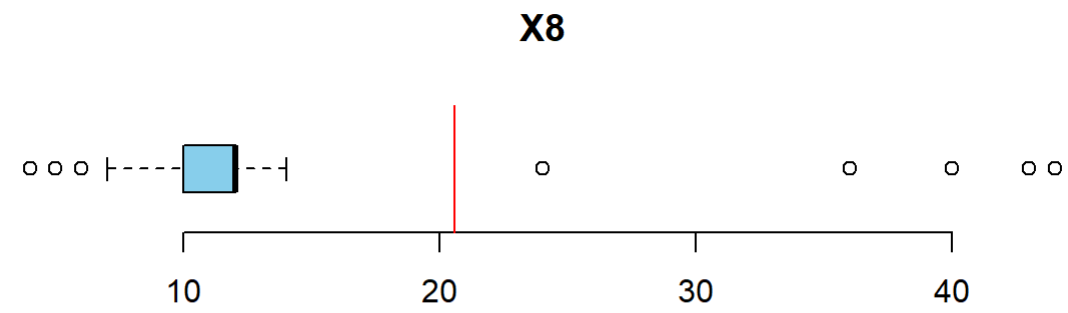
par(mfrow=c(2,1)) #Matriz de gráficos de 12x2
boxplot(col, col="skyblue", horizontal = TRUE, frame.plot=F, main = "X7")
value = quantile(col,0.75)+sd(col)
abline(v=value,col="red") #Línea vertical en el límite de los datos atípicos
```

**X7**

```
col = df$X8
# Para columna X8
cuartiles_X8 = quantile(col, c(0.25, 0.5, 0.75), type = 6);

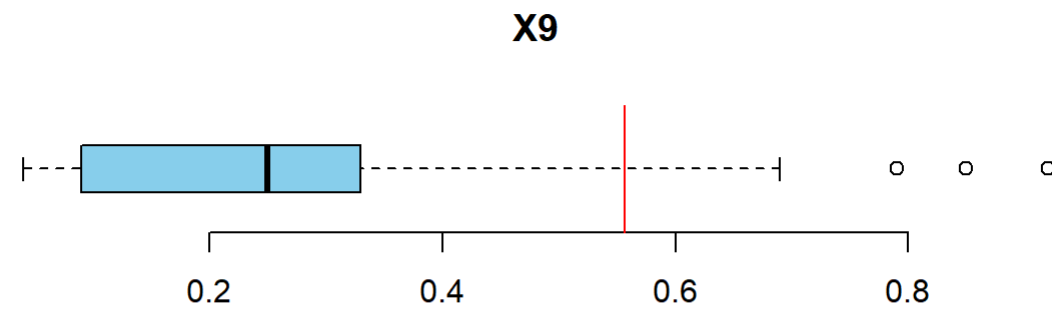
par(mfrow=c(2,1)) #Matriz de gráficos de 12x2
boxplot(col, col="skyblue", horizontal = TRUE, frame.plot=F, main = "X8")
value = quantile(col,0.75)+sd(col)
abline(v=value,col="red") #Línea vertical en el límite de los datos atípicos
quantile(df$X8, c(0.25, 0.5, 0.75), type = 6);
```





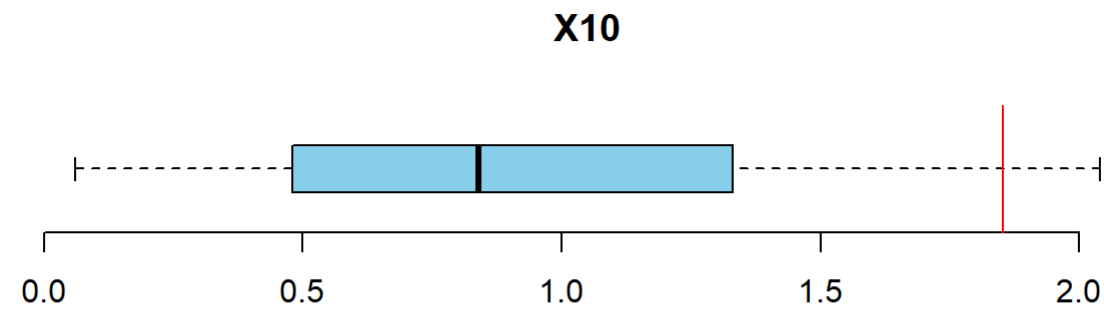
```
col = df$X9
# Para columna X9
cuartiles_X9 = quantile(col, c(0.25, 0.5, 0.75), type = 6);

par(mfrow=c(2,1)) #Matriz de gráficos de 12x2
boxplot(col, col="skyblue", horizontal = TRUE, frame.plot=F, main = "X9")
value = quantile(col,0.75)+sd(col)
abline(v=value,col="red") #Línea vertical en el límite de los datos atípicos
quantile(df$X8, c(0.25, 0.5, 0.75), type = 6);
```



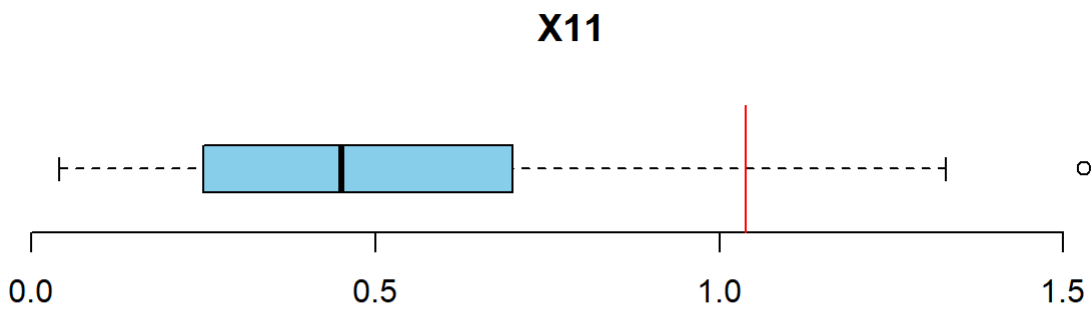
```
col = df$X10
# Para columna X10
cuartiles_X10 = quantile(col, c(0.25, 0.5, 0.75), type = 6);

par(mfrow=c(2,1)) #Matriz de gráficos de 12x2
boxplot(col, col="skyblue", horizontal = TRUE, frame.plot=F, main = "X10")
value = quantile(col,0.75)+sd(col)
abline(v=value,col="red")
```



```
col = df$X11
# Para columna X11
cuartiles_X11 = quantile(col, c(0.25, 0.5, 0.75), type = 6);

par(mfrow=c(2,1)) #Matriz de gráficos de 12x2
boxplot(col, col="skyblue", horizontal = TRUE, frame.plot=F, main = "X11")
value = quantile(col,0.75)+sd(col)
abline(v=value,col="red") #Línea vertical en el límite de los datos atípicos
quantile(df$X8, c(0.25, 0.5, 0.75), type = 6);
```

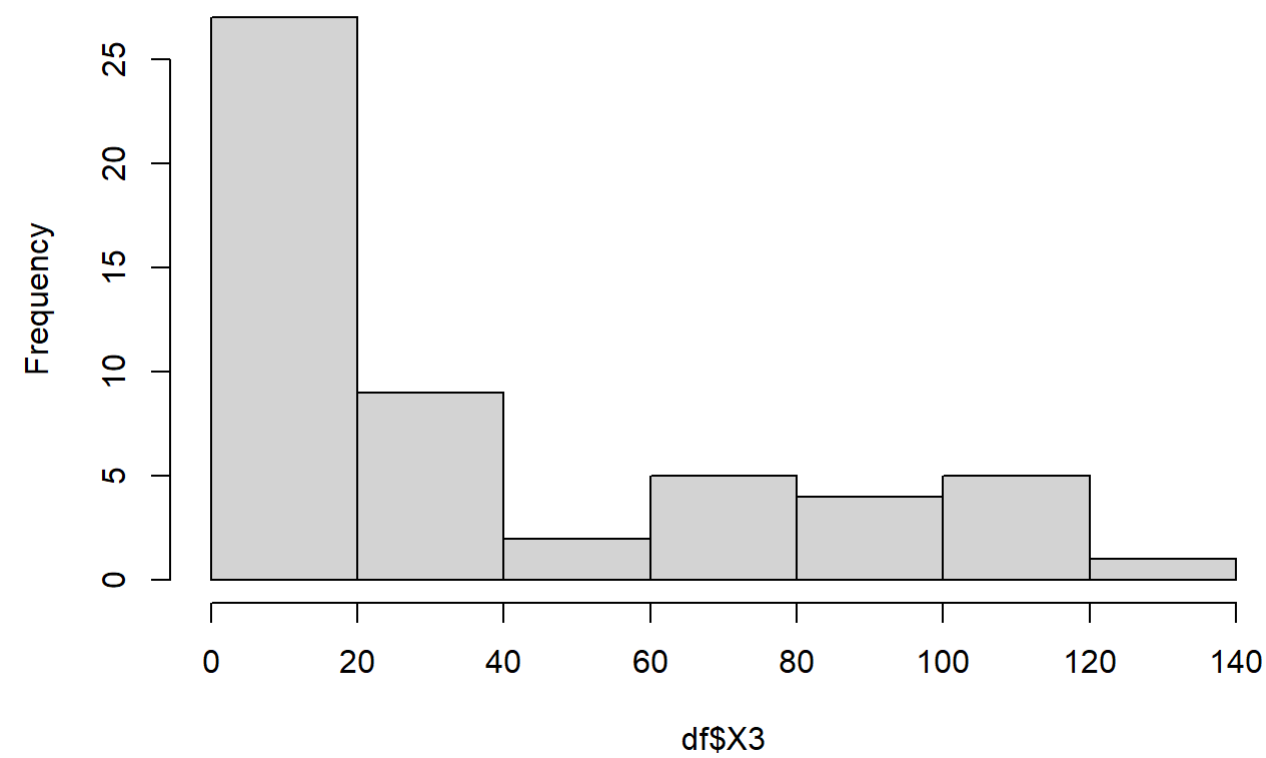


####Análisis de distribución de los

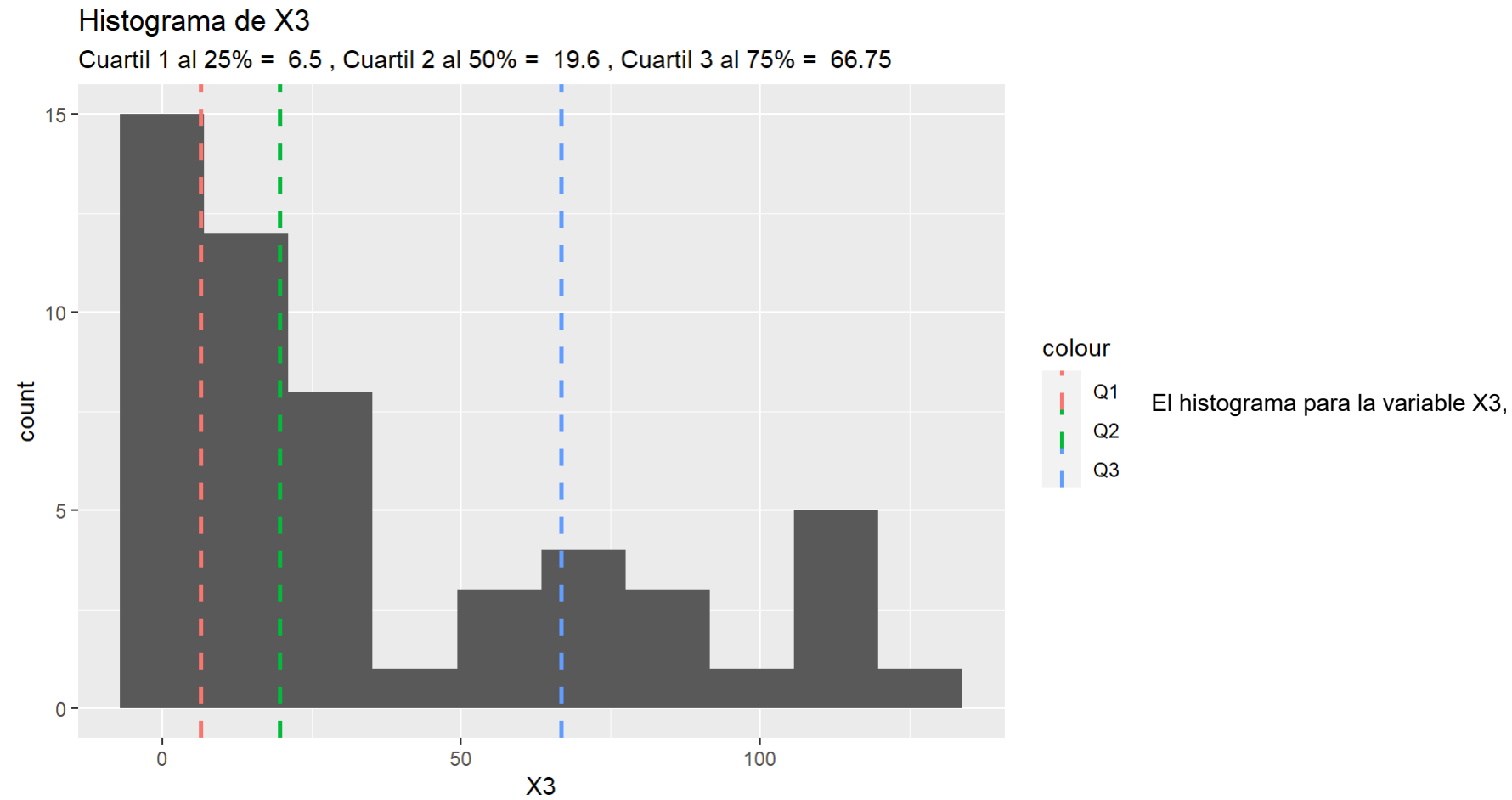
datos (Histogramas). Identificar si tiene forma simétrica o asimétrica

```
hist(df$X3)
Q1 = cuartiles_X3[1]
Q2 = cuartiles_X3[2]
Q3 = cuartiles_X3[3]
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

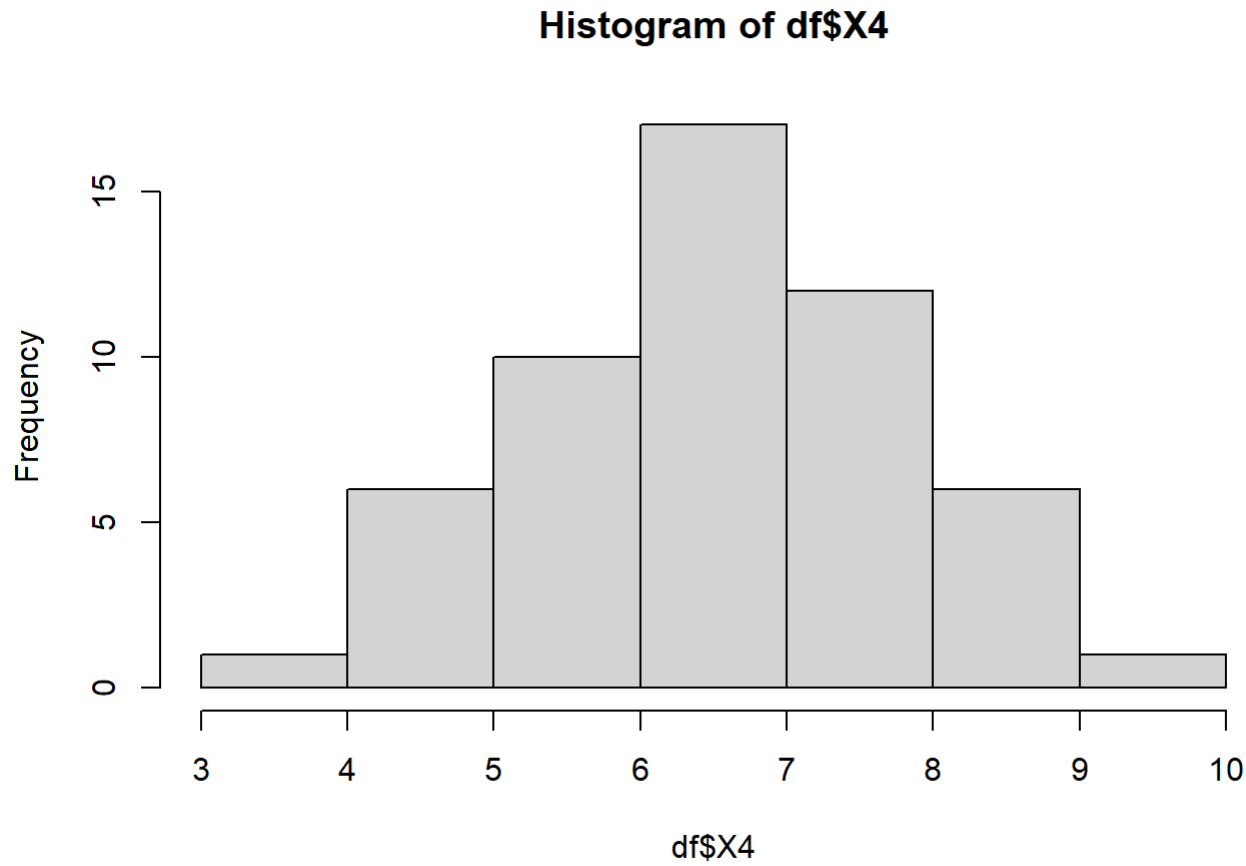
**Histogram of df\$X3**

```
ggplot(data = df, aes(x=X3)) +  
  geom_histogram(bins = 10) +  
  geom_vline(aes(xintercept = Q1,  
    color = "Q1"),  
    linetype = "dashed",  
    size = 1) +  
  geom_vline(aes(xintercept = Q2,  
    color = "Q2"),  
    linetype = "dashed",  
    size = 1) +  
  geom_vline(aes(xintercept = Q3,  
    color = "Q3"),  
    linetype = "dashed",  
    size = 1) +  
  labs(title = "Histograma de X3", subtitle = paste("Cuartil 1 al 25% = ",Q1, ", Cuartil 2 al 50% = ",Q2, ", Cuartil 3 al 75% = ",Q3))
```



tiene forma asimétrica

```
hist(df$X4)
```



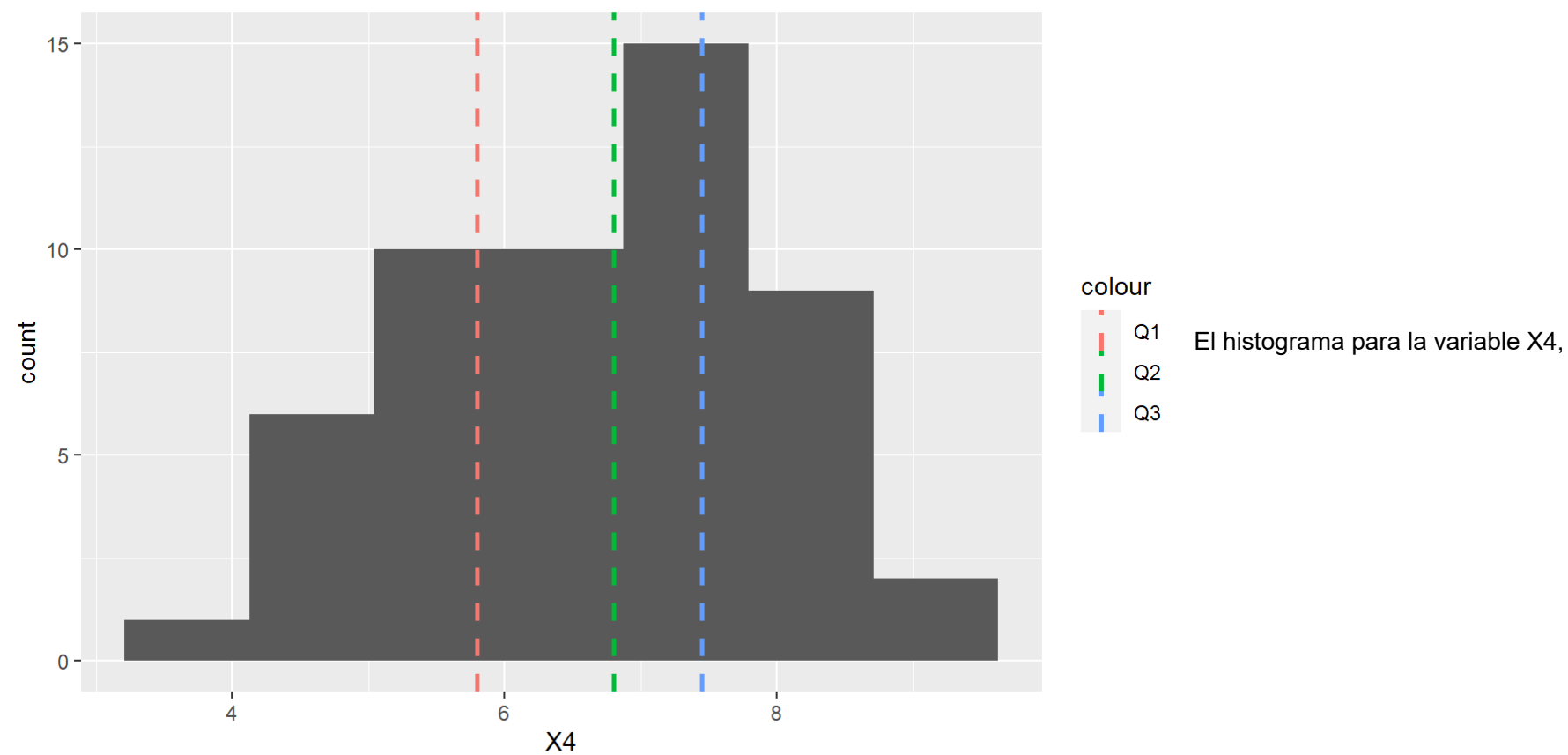
```

Q1 = cuartiles_X4[1]
Q2 = cuartiles_X4[2]
Q3 = cuartiles_X4[3]
ggplot(data = df, aes(x=X4)) +
  geom_histogram(bins = 7) +
  geom_vline(aes(xintercept = Q1,
                 color = "Q1"),
             linetype = "dashed",
             size = 1) +
  geom_vline(aes(xintercept = Q2,
                 color = "Q2"),
             linetype = "dashed",
             size = 1) +
  geom_vline(aes(xintercept = Q3,
                 color = "Q3"),
             linetype = "dashed",
             size = 1) +
  labs(title = "Histograma de X4", subtitle = paste("Cuartil 1 al 25% = ",Q1, ", Cuartil 2 al 50% = ",Q2, ", Cuartil 3 al 75% = ",Q3))

```

### Histograma de X4

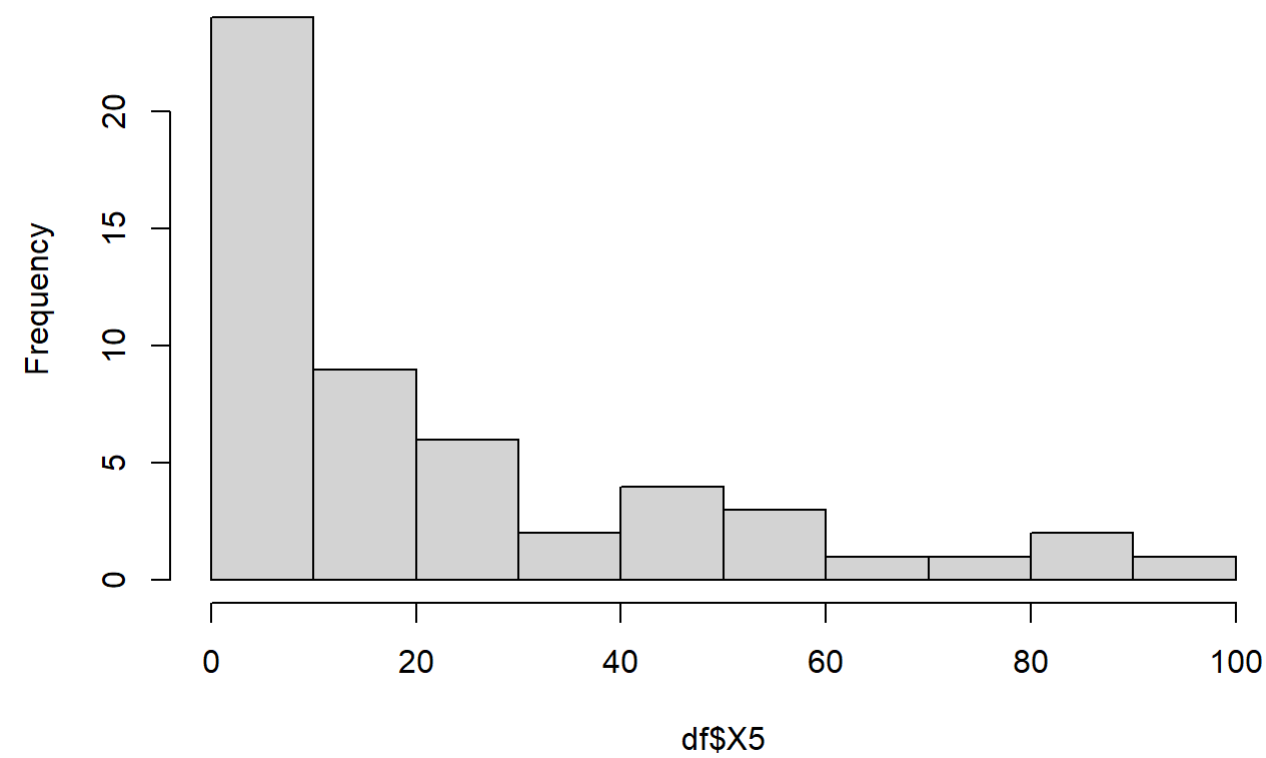
Cuartil 1 al 25% = 5.8 , Cuartil 2 al 50% = 6.8 , Cuartil 3 al 75% = 7.45



tiene forma asimétrica

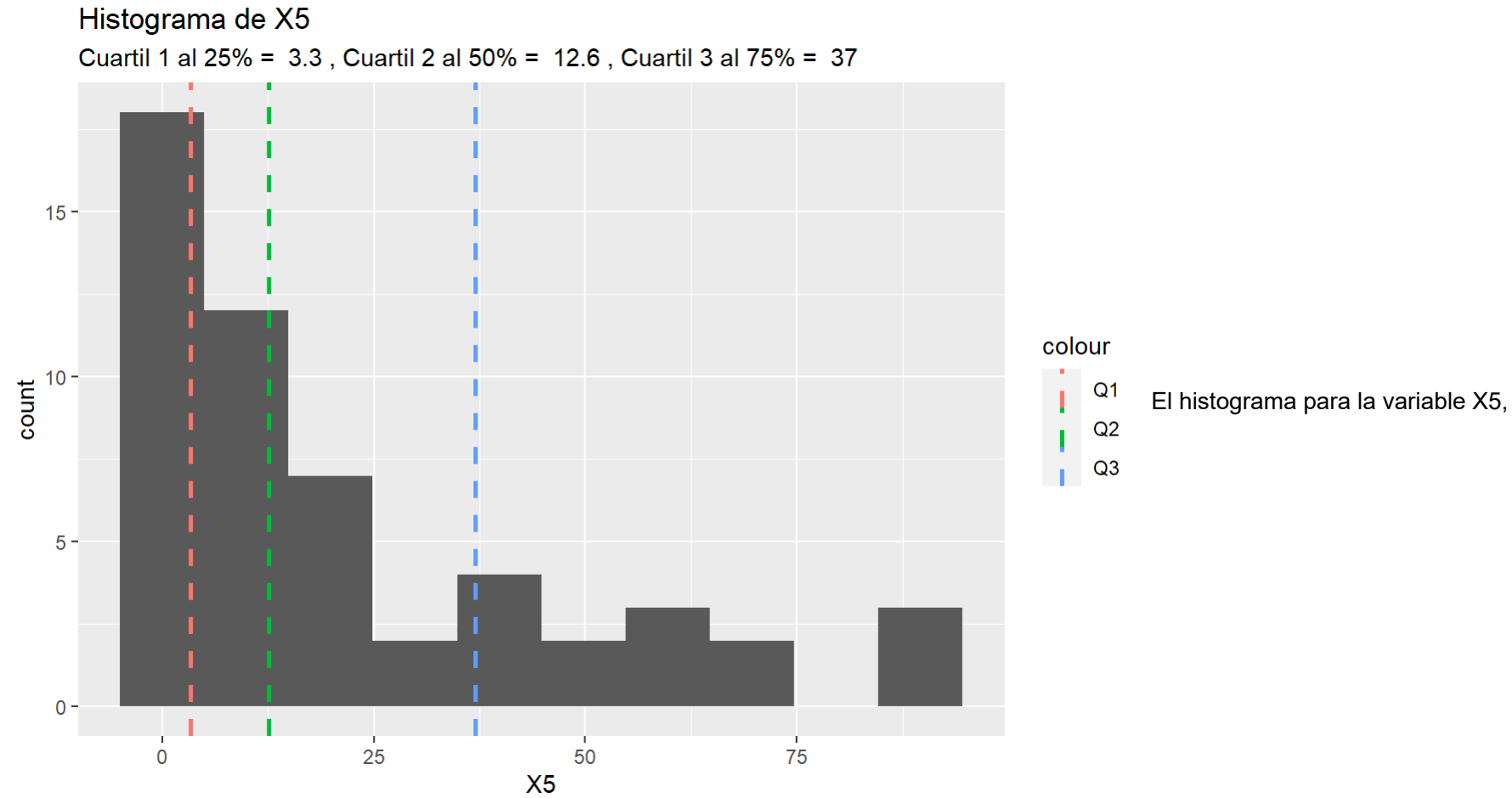
```
hist(df$X5)
```

## Histogram of df\$X5



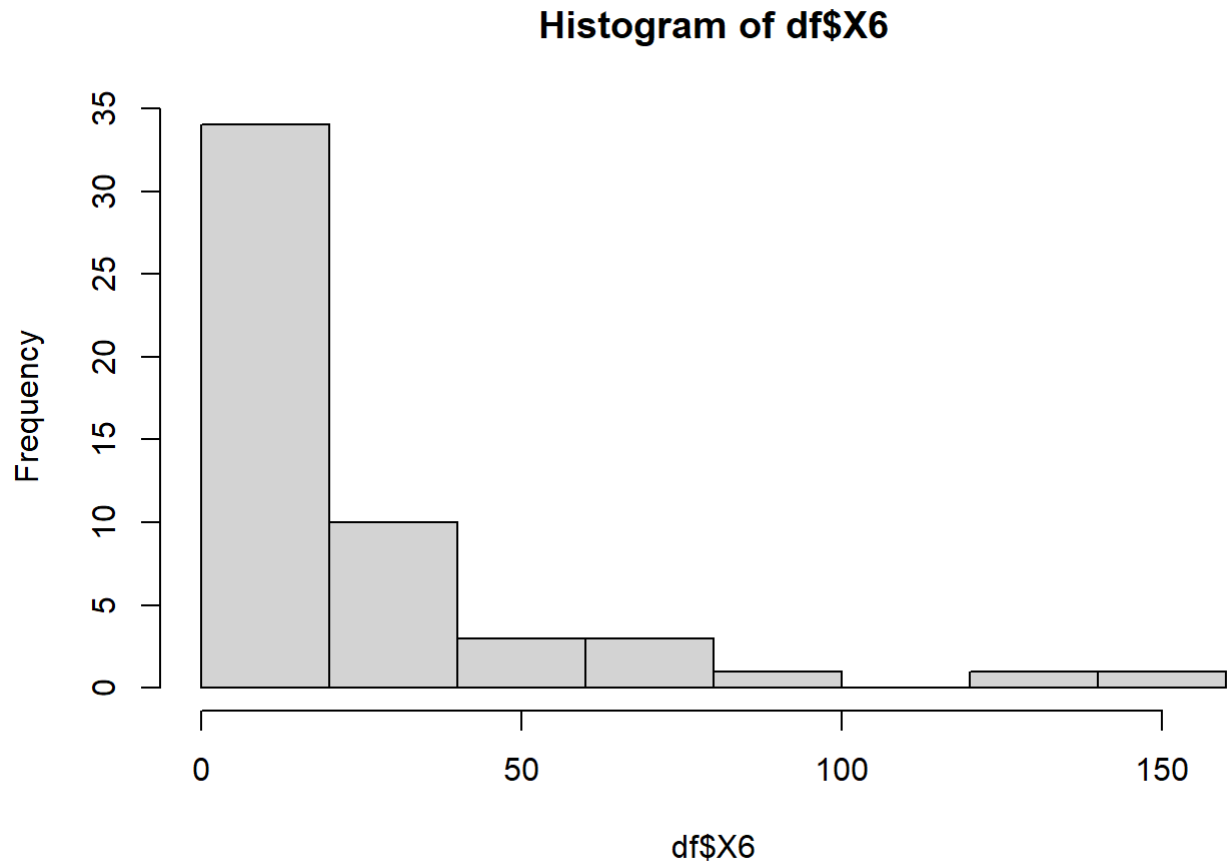
```
Q1 = cuartiles_X5[1]
Q2 = cuartiles_X5[2]
Q3 = cuartiles_X5[3]
ggplot(data = df, aes(x=X5)) +
  geom_histogram(bins = 10) +
  geom_vline(aes(xintercept = Q1,
                 color = "Q1"),
             linetype = "dashed",
             size = 1) +
  geom_vline(aes(xintercept = Q2,
                 color = "Q2"),
             linetype = "dashed",
             size = 1) +
  geom_vline(aes(xintercept = Q3,
                 color = "Q3"),
             linetype = "dashed",
             size = 1) +
  labs(title = "Histograma de X5", subtitle = paste("Cuartil 1 al 25% = ",Q1, ", ", Cuartil 2 al 50% = ",Q2, ", ", Cuartil 3 al 75% = ",Q3))
```





tiene forma asimétrica

```
hist(df$X6)
```



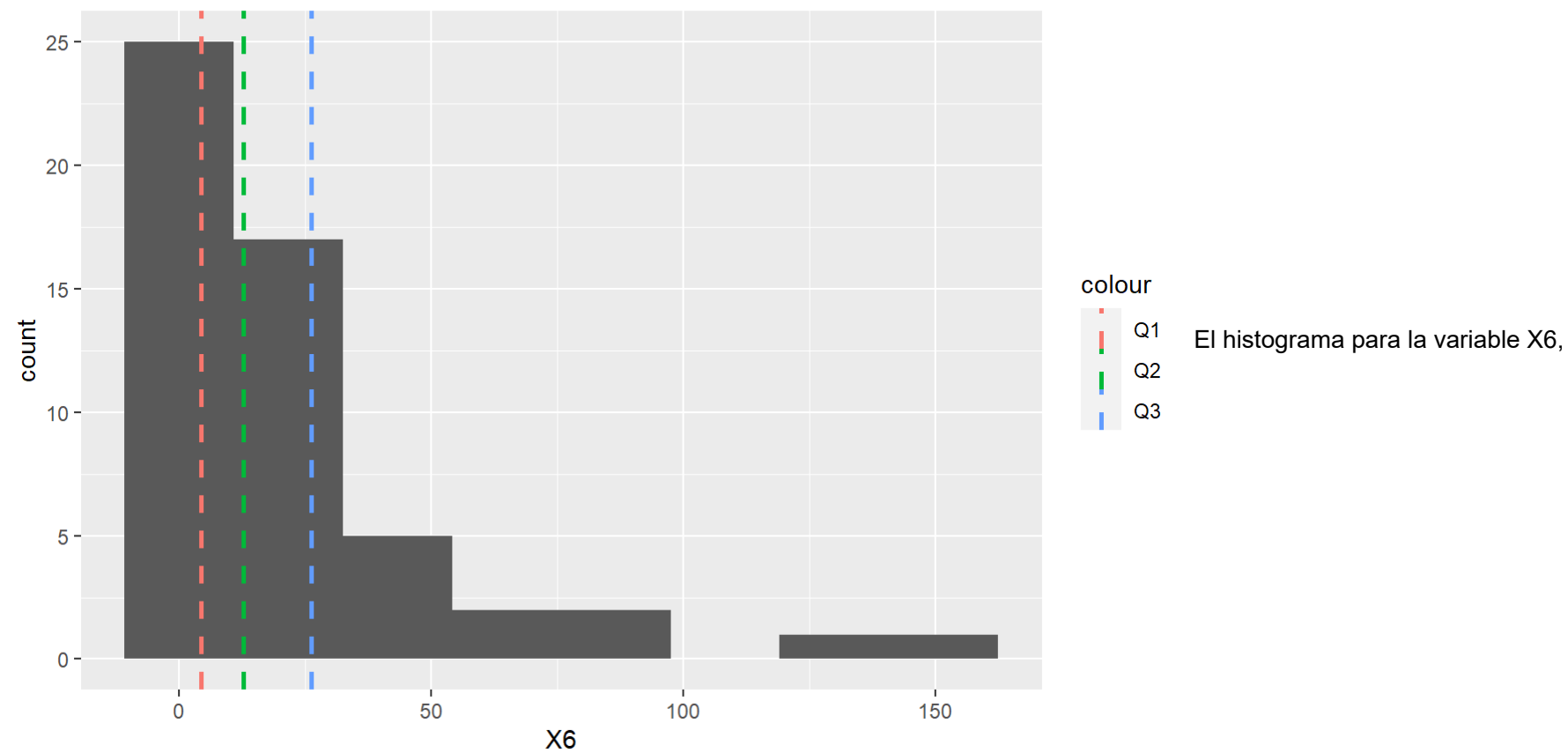
```

Q1 = cuartiles_X6[1]
Q2 = cuartiles_X6[2]
Q3 = cuartiles_X6[3]
ggplot(data = df, aes(x=X6)) +
  geom_histogram(bins = 8) +
  geom_vline(aes(xintercept = Q1,
                 color = "Q1"),
             linetype = "dashed",
             size = 1) +
  geom_vline(aes(xintercept = Q2,
                 color = "Q2"),
             linetype = "dashed",
             size = 1) +
  geom_vline(aes(xintercept = Q3,
                 color = "Q3"),
             linetype = "dashed",
             size = 1) +
  labs(title = "Histograma de X6", subtitle = paste("Cuartil 1 al 25% = ",Q1, ", Cuartil 2 al 50% = ",Q2, ", Cuartil 3 al 75% = ",Q3))

```

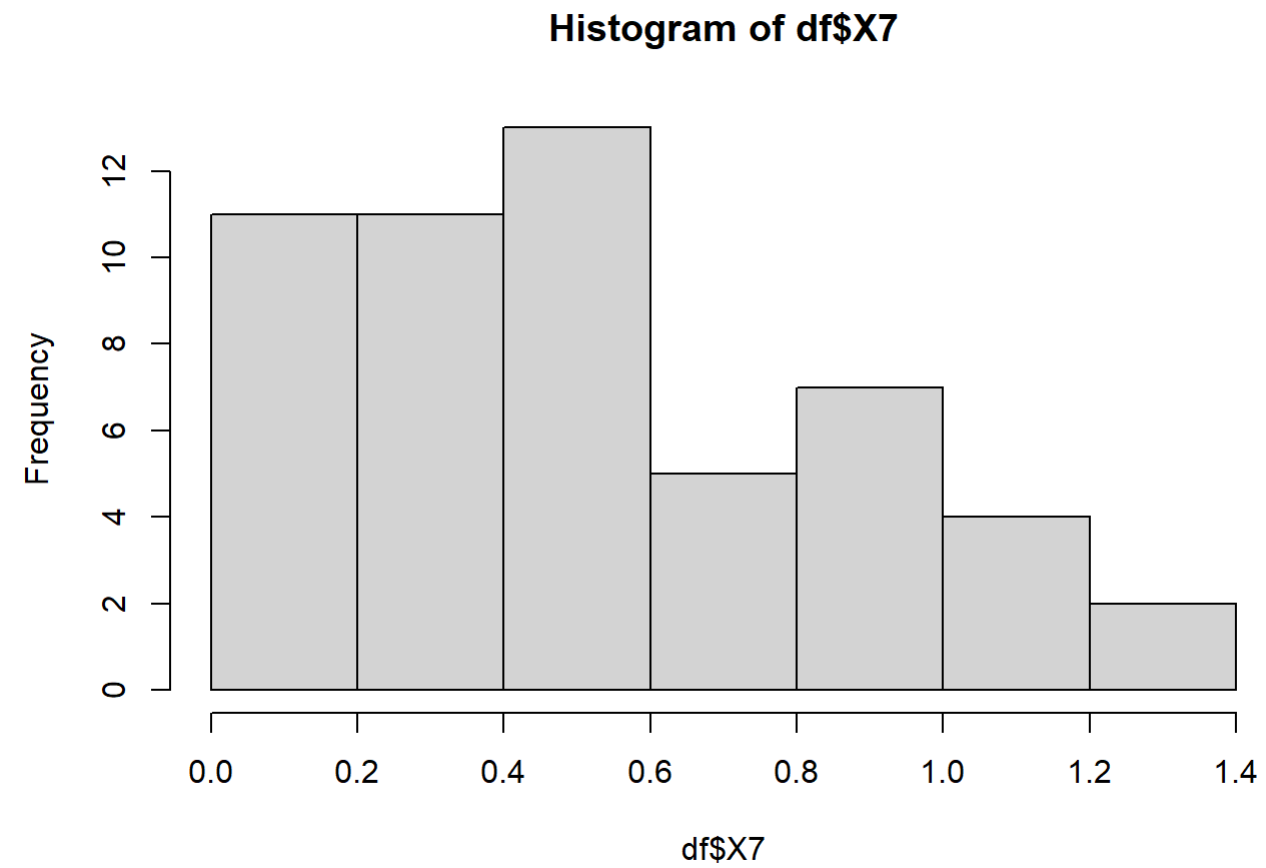
### Histograma de X6

Cuartil 1 al 25% = 4.3 , Cuartil 2 al 50% = 12.8 , Cuartil 3 al 75% = 26.2

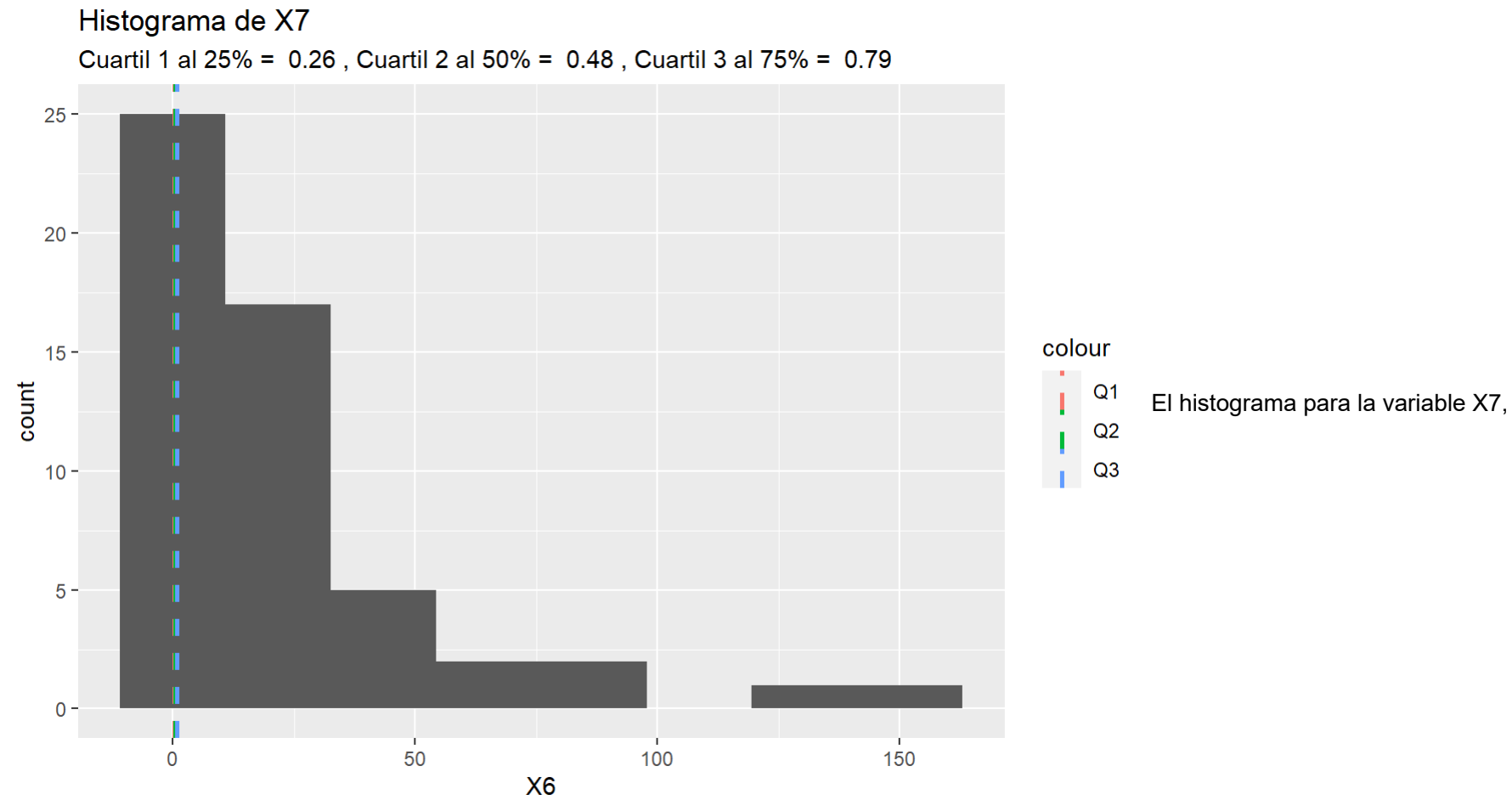


tiene forma asimétrica

```
hist(df$X7)
```

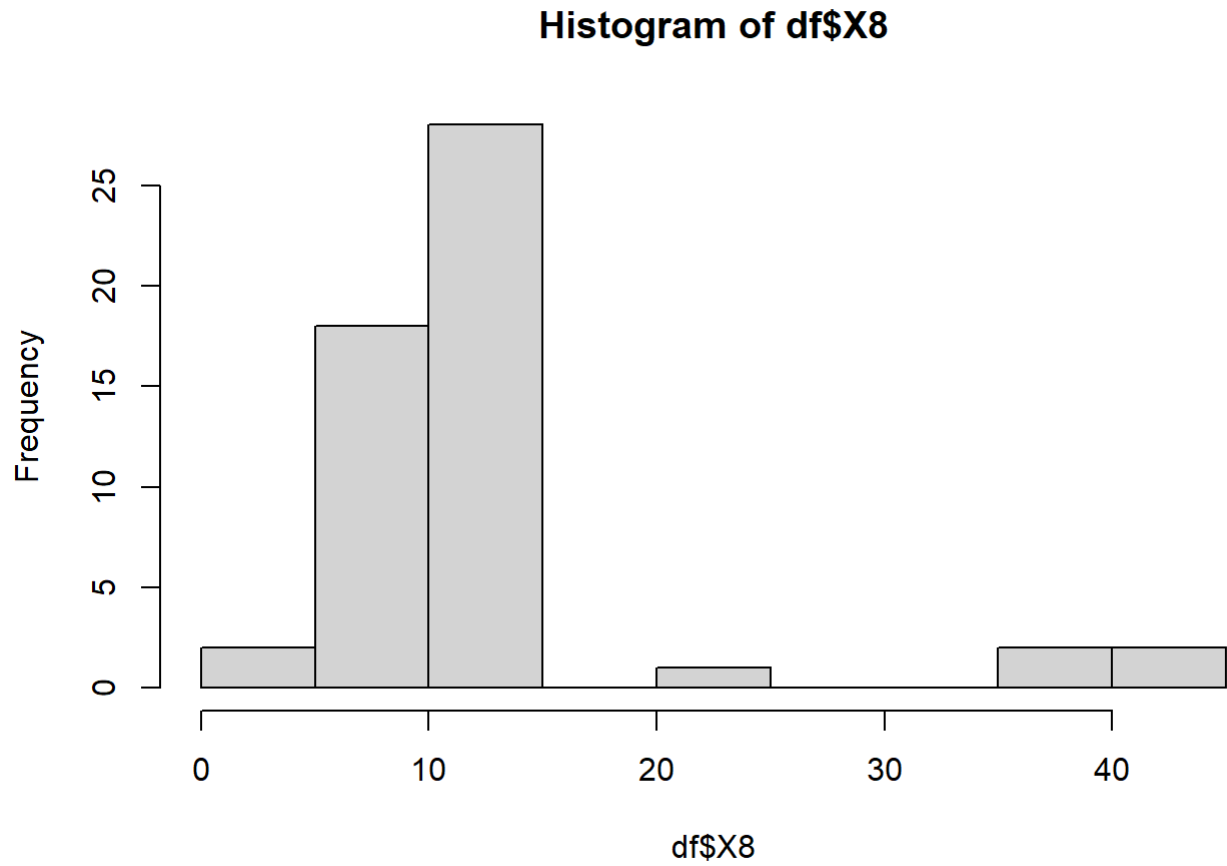


```
Q1 = cuartiles_X7[1]
Q2 = cuartiles_X7[2]
Q3 = cuartiles_X7[3]
ggplot(data = df, aes(x=X6)) +
  geom_histogram(bins = 8) +
  geom_vline(aes(xintercept = Q1,
                 color = "Q1"),
             linetype = "dashed",
             size = 1) +
  geom_vline(aes(xintercept = Q2,
                 color = "Q2"),
             linetype = "dashed",
             size = 1) +
  geom_vline(aes(xintercept = Q3,
                 color = "Q3"),
             linetype = "dashed",
             size = 1) +
  labs(title = "Histograma de X7", subtitle = paste("Cuartil 1 al 25% = ",Q1, ", Cuartil 2 al 50% = ",Q2, ", Cuartil 3 al 75% = ",Q3))
```



tiene forma asimétrica

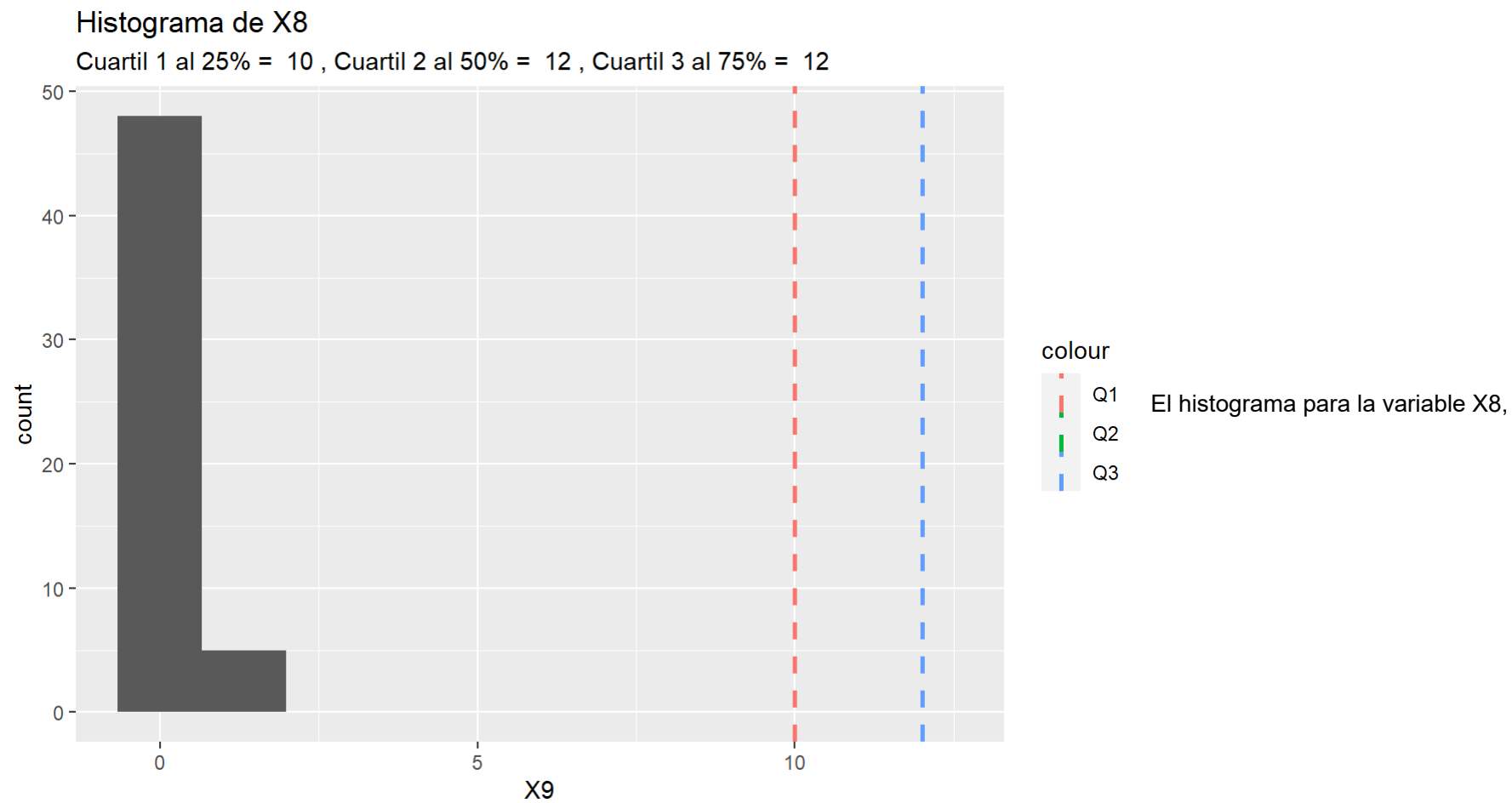
```
hist(df$X8)
```



```

Q1 = cuartiles_X8[1]
Q2 = cuartiles_X8[2]
Q3 = cuartiles_X8[3]
ggplot(data = df, aes(x=X9)) +
  geom_histogram(bins = 10) +
  geom_vline(aes(xintercept = Q1,
                 color = "Q1"),
             linetype = "dashed",
             size = 1) +
  geom_vline(aes(xintercept = Q2,
                 color = "Q2"),
             linetype = "dashed",
             size = 1) +
  geom_vline(aes(xintercept = Q3,
                 color = "Q3"),
             linetype = "dashed",
             size = 1) +
  labs(title = "Histograma de X8", subtitle = paste("Cuartil 1 al 25% = ",Q1, ", Cuartil 2 al 50% = ",Q2, ", Cuartil 3 al 75% = ",Q3))

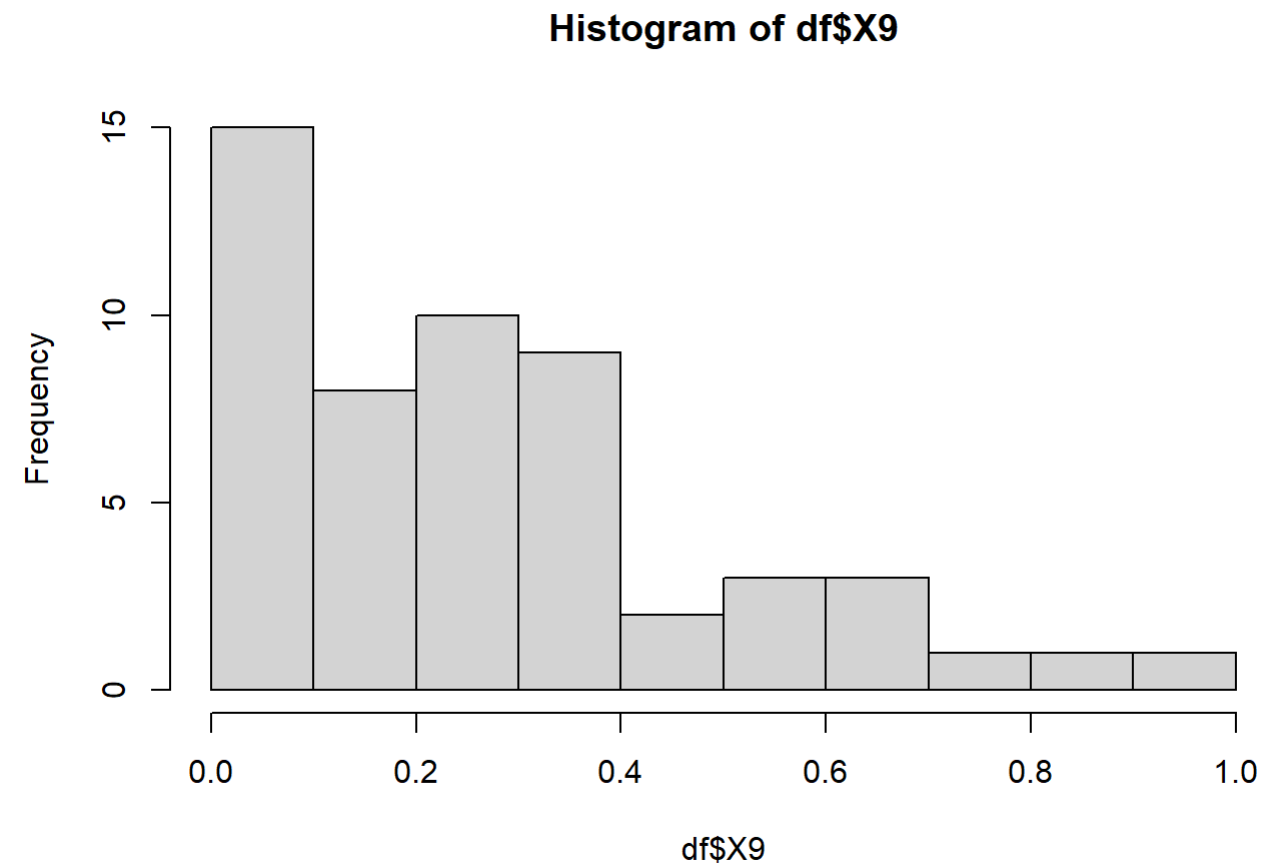
```



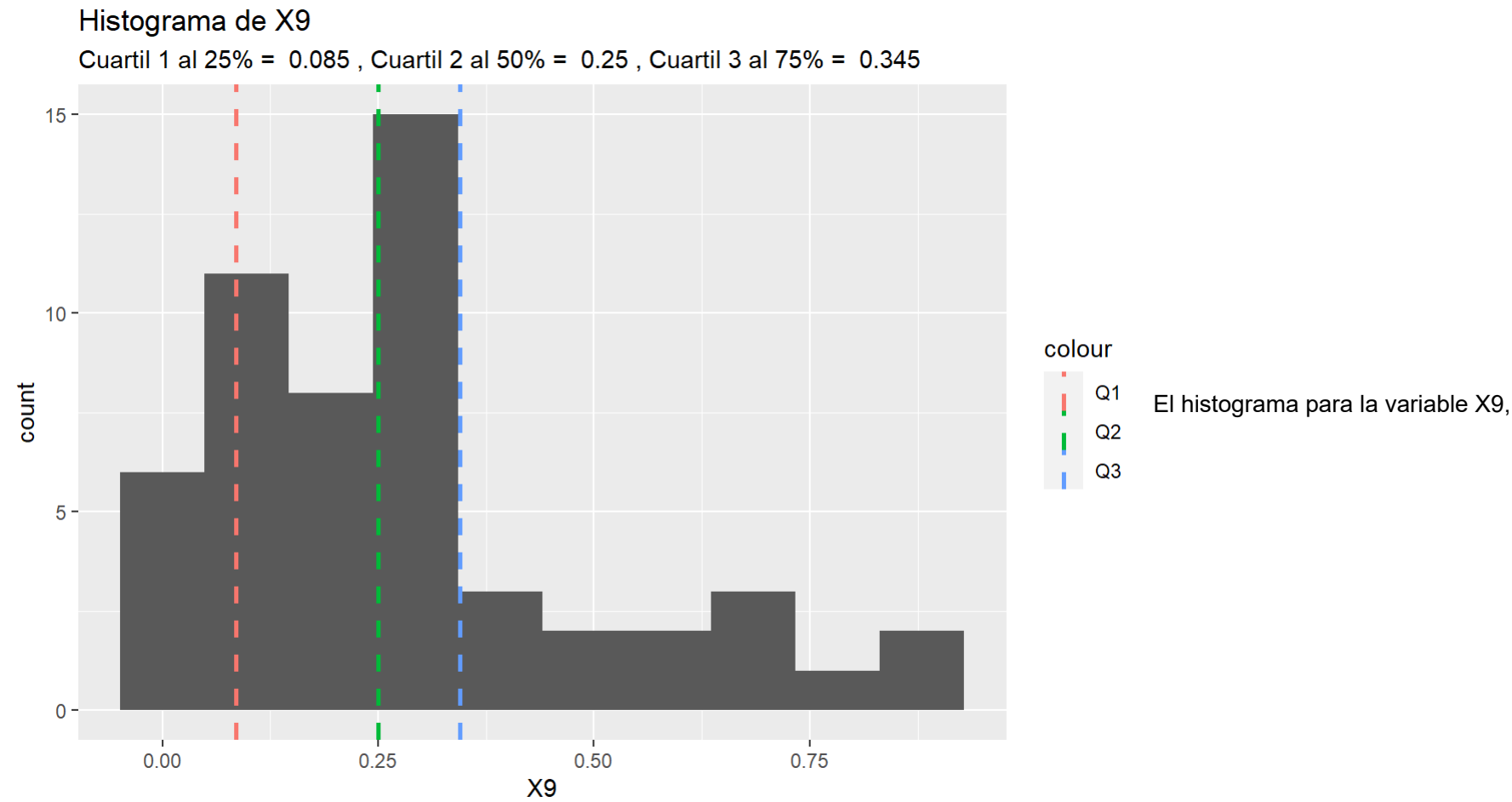
El histograma para la variable X8,

tiene forma asimétrica

```
hist(df$X9)
```

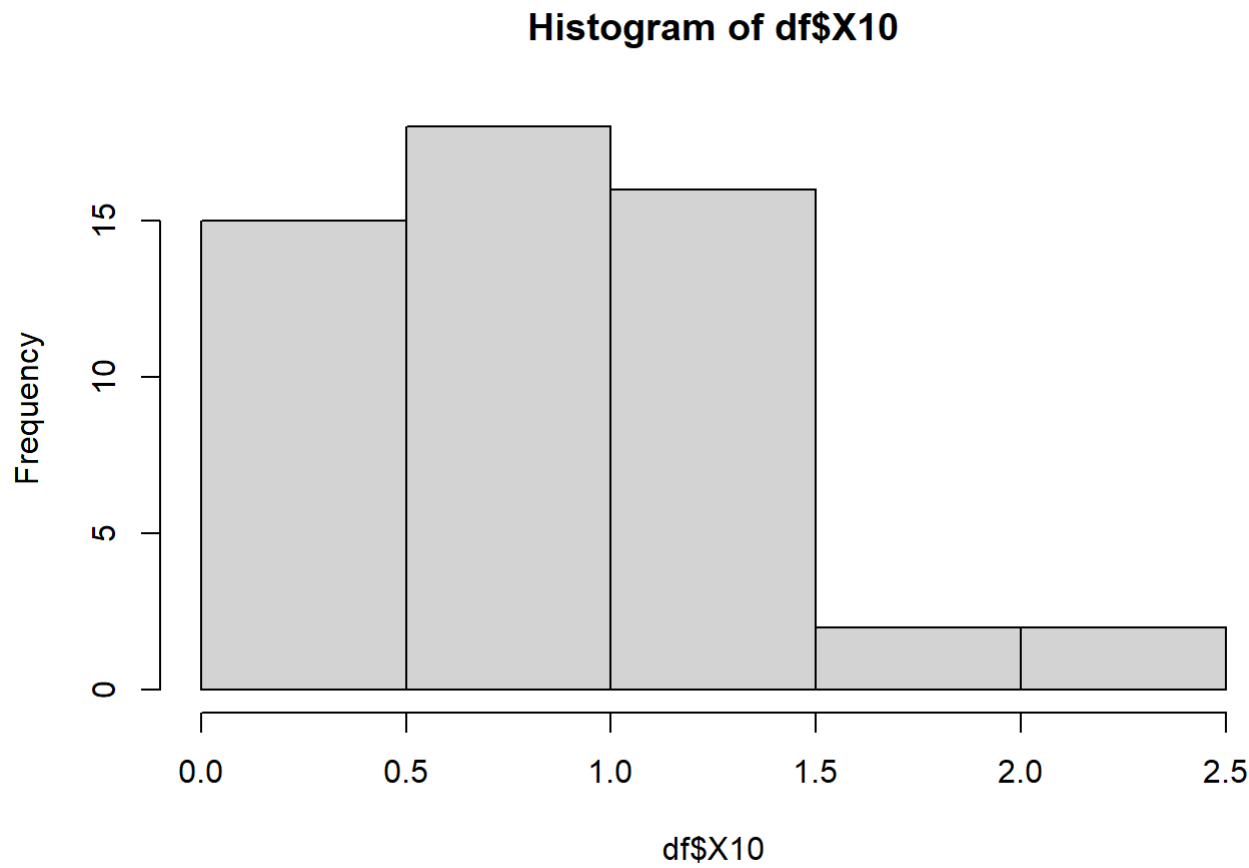


```
Q1 = cuartiles_X9[1]
Q2 = cuartiles_X9[2]
Q3 = cuartiles_X9[3]
ggplot(data = df, aes(x=X9)) +
  geom_histogram(bins = 10) +
  geom_vline(aes(xintercept = Q1,
                 color = "Q1"),
            linetype = "dashed",
            size = 1) +
  geom_vline(aes(xintercept = Q2,
                 color = "Q2"),
            linetype = "dashed",
            size = 1) +
  geom_vline(aes(xintercept = Q3,
                 color = "Q3"),
            linetype = "dashed",
            size = 1) +
  labs(title = "Histograma de X9", subtitle = paste("Cuartil 1 al 25% = ",Q1, ", Cuartil 2 al 50% = ",Q2, ", Cuartil 3 al 75% = ",Q3))
```



tiene forma asimétrica

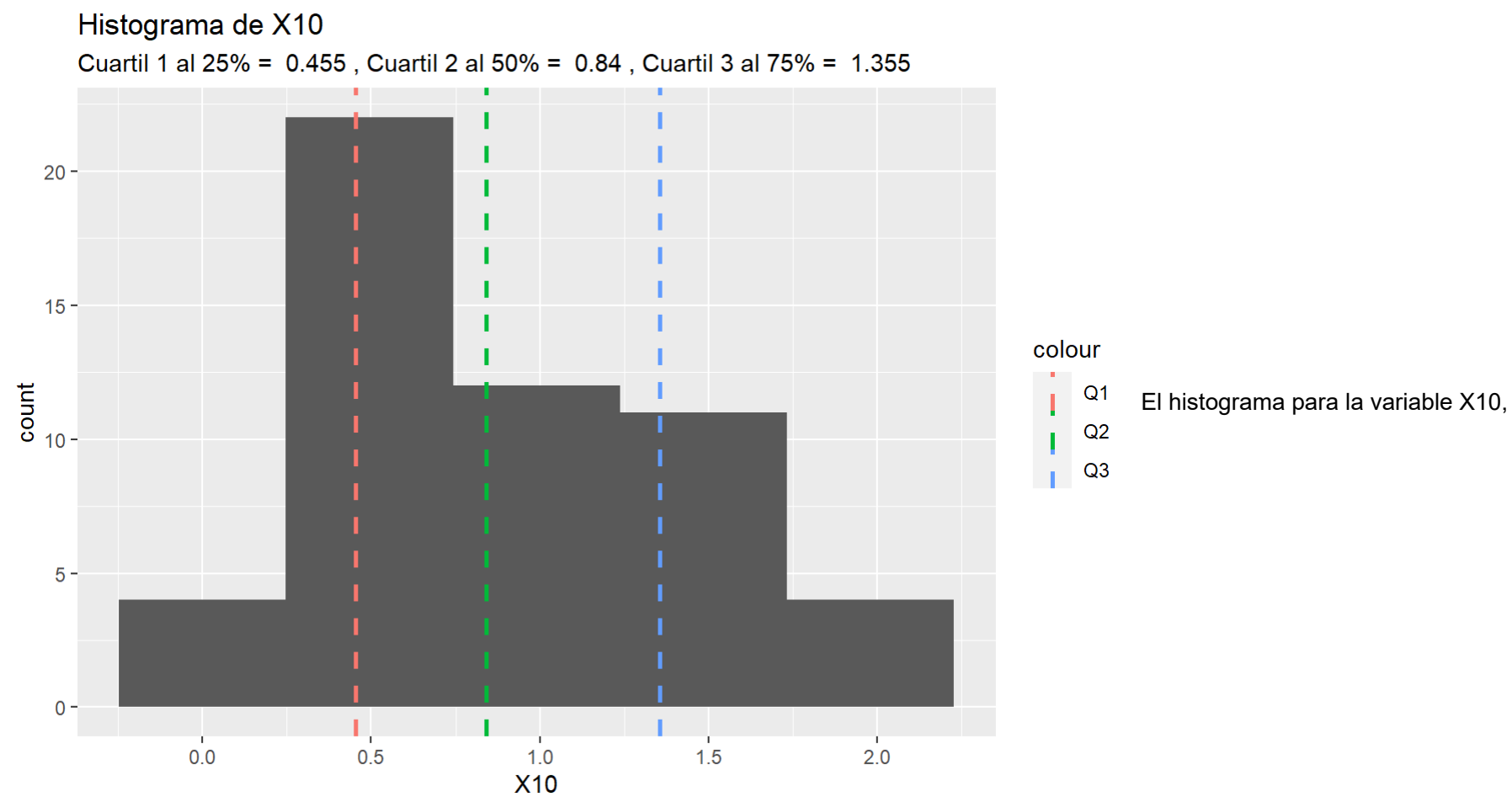
```
hist(df$X10)
```



```

Q1 = cuartiles_X10[1]
Q2 = cuartiles_X10[2]
Q3 = cuartiles_X10[3]
ggplot(data = df, aes(x=X10)) +
  geom_histogram(bins = 5) +
  geom_vline(aes(xintercept = Q1,
                 color = "Q1",
                 linetype = "dashed",
                 size = 1) +
  geom_vline(aes(xintercept = Q2,
                 color = "Q2",
                 linetype = "dashed",
                 size = 1) +
  geom_vline(aes(xintercept = Q3,
                 color = "Q3",
                 linetype = "dashed",
                 size = 1) +
  labs(title = "Histograma de X10", subtitle = paste("Cuartil 1 al 25% = ", Q1, ", Cuartil 2 al 50% = ", Q2, ", Cuartil 3 al 75% = ", Q3))

```

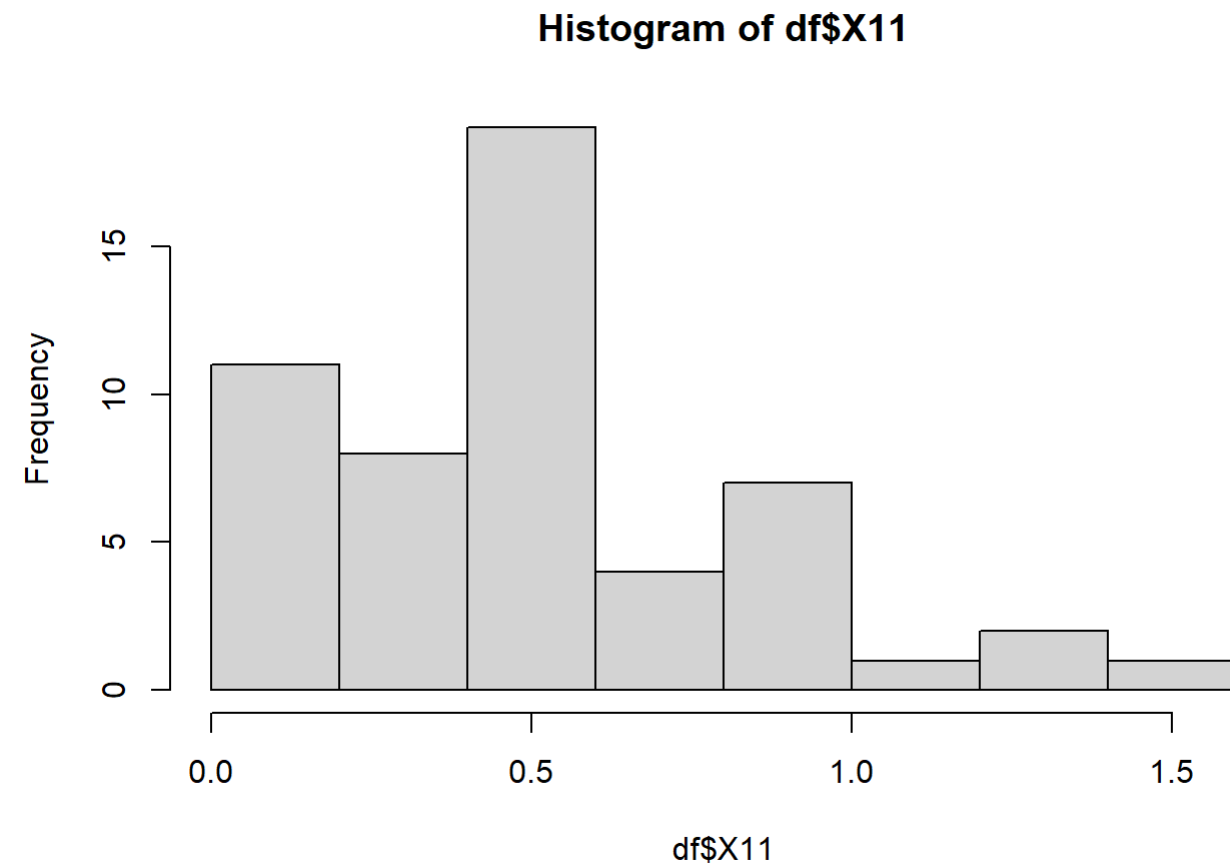


El histograma para la variable X10,

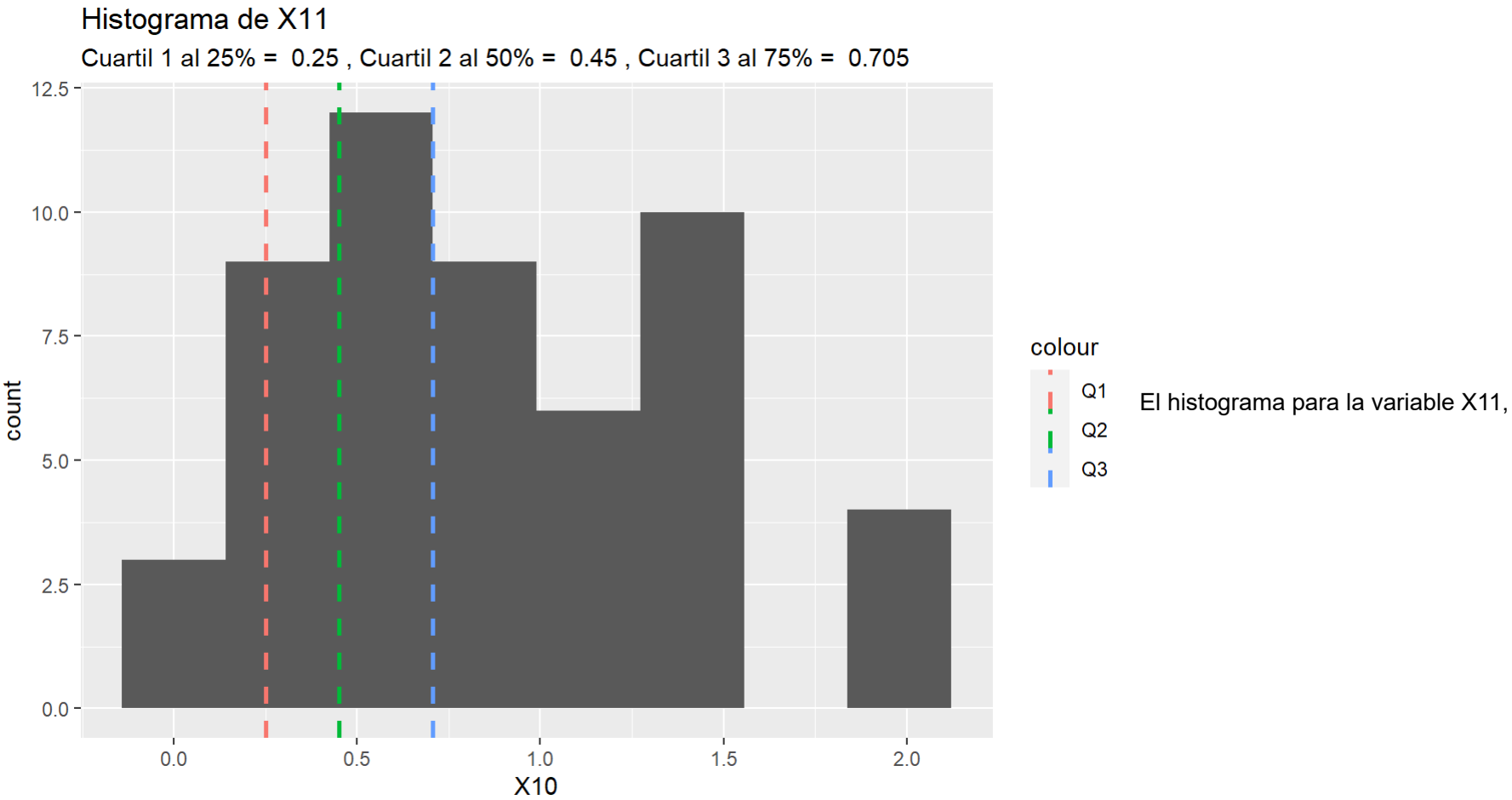
tiene forma asimétrica

```
hist(df$X11)
```





```
Q1 = cuartiles_X11[1]
Q2 = cuartiles_X11[2]
Q3 = cuartiles_X11[3]
ggplot(data = df, aes(x=X10)) +
  geom_histogram(bins = 8) +
  geom_vline(aes(xintercept = Q1,
                 color = "Q1"),
             linetype = "dashed",
             size = 1) +
  geom_vline(aes(xintercept = Q2,
                 color = "Q2"),
             linetype = "dashed",
             size = 1) +
  geom_vline(aes(xintercept = Q3,
                 color = "Q3"),
             linetype = "dashed",
             size = 1) +
  labs(title = "Histograma de X11", subtitle = paste("Cuartil 1 al 25% = ",Q1, ", Cuartil 2 al 50% = ",Q2, ", Cuartil 3 al 75% = ",Q3))
```



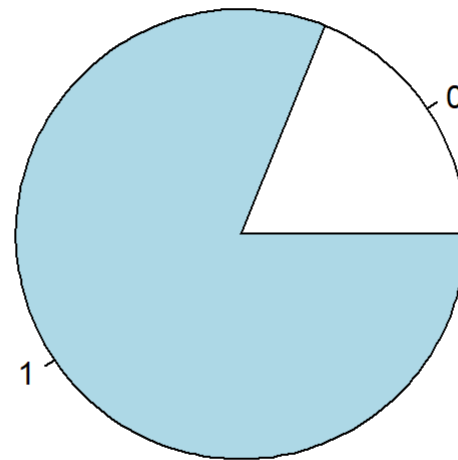
tiene forma asimétrica

###Variables categóricas ####Distribución de los datos (diagramas de barras, diagramas de pastel)

```
# Tabla de frecuencias
mi_tabla <- table(df$X12)

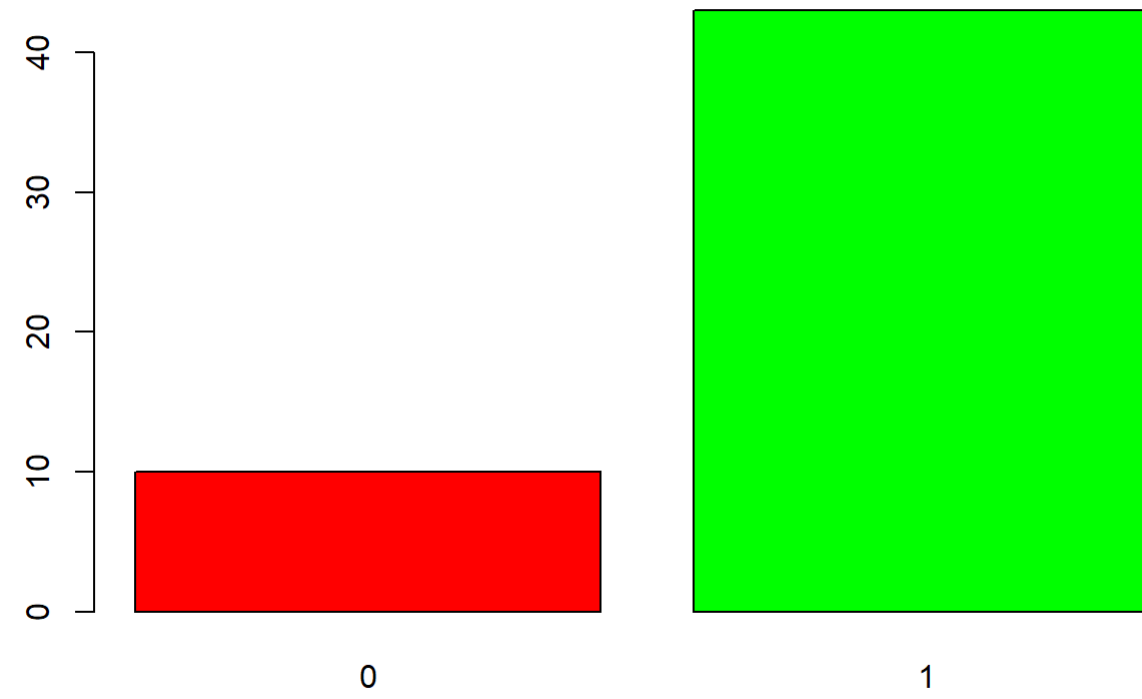
pie(mi_tabla, main ="Diagrama de Pastel para Edad de Peces")
```

## Diagrama de Pastel para Edad de Peces



```
# Gráfico de barras de frecuencia absoluta  
barplot(mi_tabla, main = "Frecuencia absoluta Edad Peces",  
        col = rainbow(3))
```

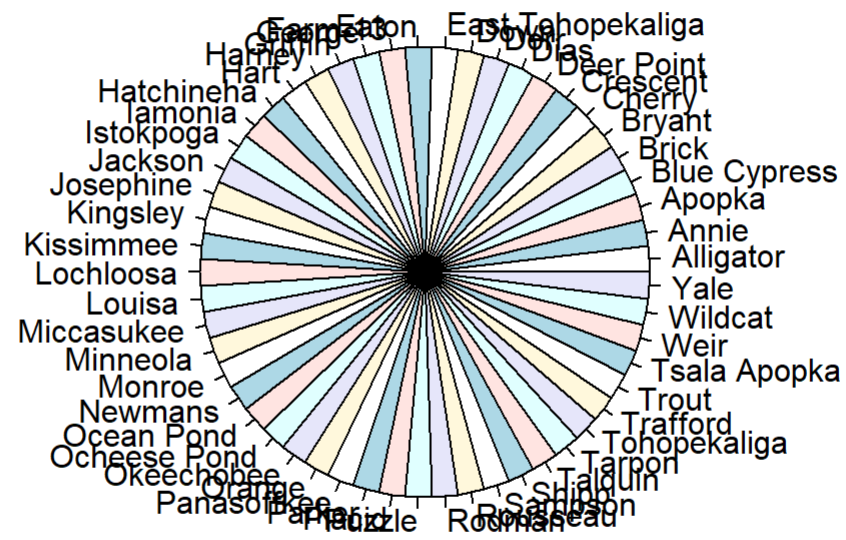
### Frecuencia absoluta Edad Peces



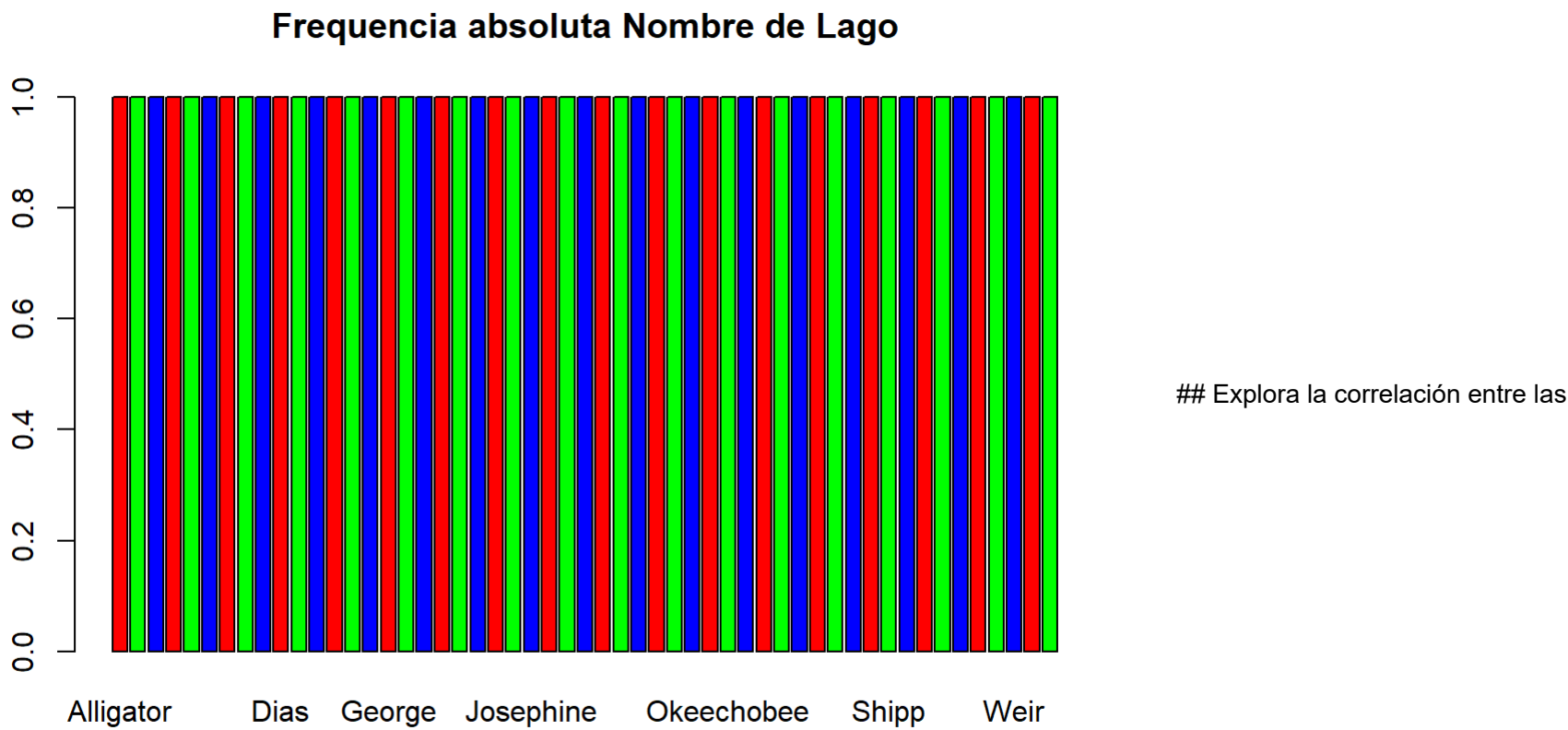
```
tabla_edad <- table(df$X2)

pie(tabla_edad, main = "Diagrama de Pastel para Nombre de Lago")
```

### Diagrama de Pastel para Nombre de Lago



```
# Gráfico de barras de frecuencia absoluta
barplot(tabla_edad, main = "Frecuencia absoluta Nombre de Lago",
        col = rainbow(3))
```



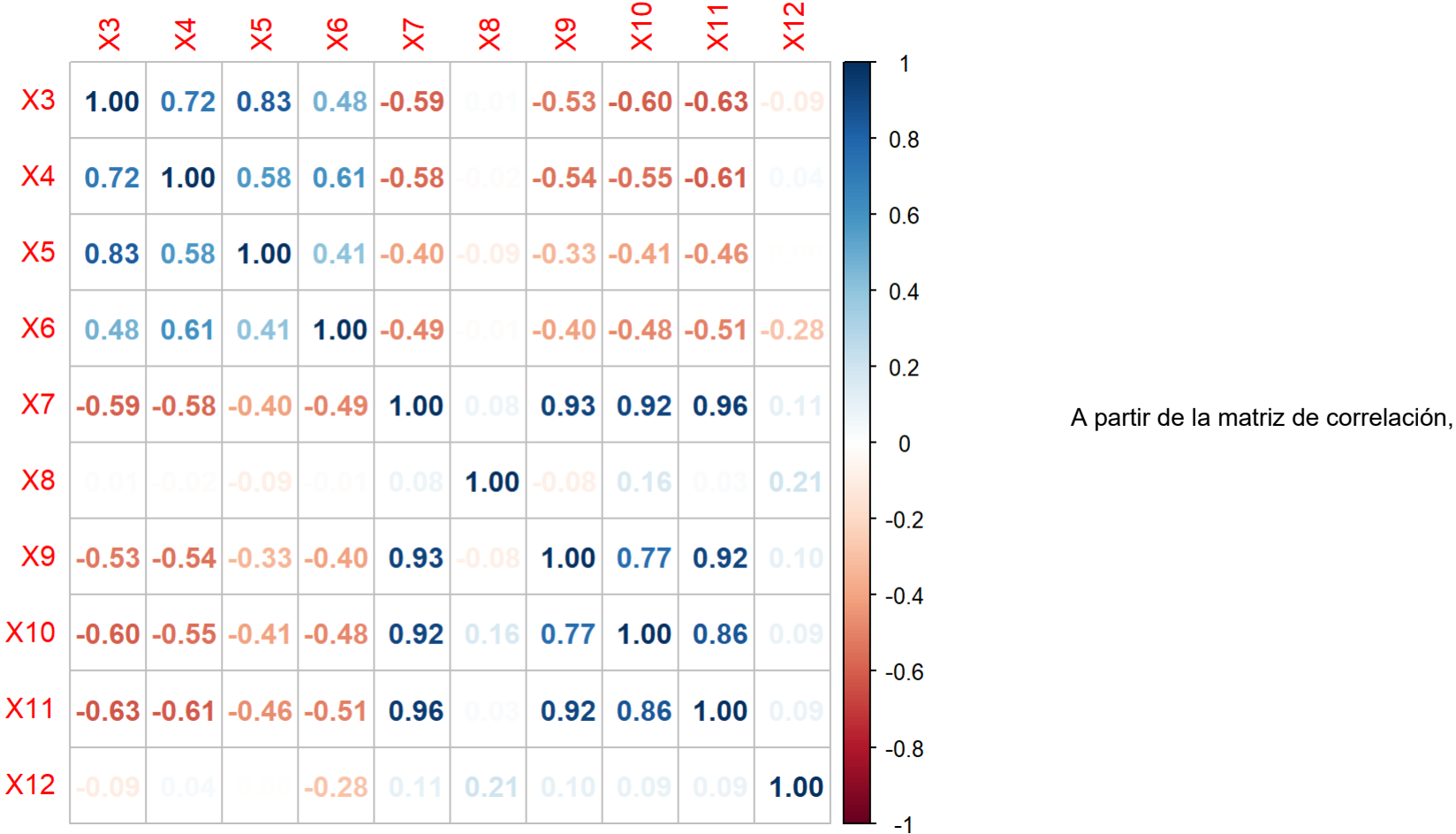
variables. Identifica cuáles son las correlaciones más fuertes y qué sentido tiene relacionarlas.

```
#install.packages("corrplot")
library(corrplot)

## Warning: package 'corrplot' was built under R version 4.1.3

## corrplot 0.92 loaded

C = cor(df[,c(-1,-2)])
#Se quitan las primeras dos variables del análisis ya que únicamente tienen propósitos de identificación de cada dato.
corrplot(C, method = 'number')
```



A partir de la matriz de correlación,

se puede ver que las variables que tienen las correlaciones más fuertes corresponden a: \* X11 y X7: 0.96 \* X9 y X7: 0.93 \* X10 y X7: 0.92 \* X11 y X10: 0.86 \* X3 y X5: 0.83 \* X10 y X9: 0.77 \* X3 y X11: -0.63 \* X4 y X11: -0.61 \* X3 y X10: -0.60

X3 = alcalinidad (mg/l de carbonato de calcio) X4 = pH X5 = calcio (mg/l) X7 = concentración media de mercurio (parte por millón) en el tejido muscular del grupo de peces estudiados en cada lago X9 = mínimo de la concentración de mercurio en cada grupo de peces X10 = máximo de la concentración de mercurio en cada grupo de peces X11 = estimación (mediante regresión) de la concentración de mercurio en el pez de 3 años (o promedio de mercurio cuando la edad no está disponible)

Tiene sentido analizar la relación entre la cantidad de calcio (X5) con la alcalinidad (X3). Tiene sentido analizar la relación entre el nivel de alcalinidad (X3) con la concentración de mercurio en el pez de 3 años (X11) Tiene sentido analizar la relación entre el nivel de pH (X4) con la concentración de mercurio en el pez de 3 años (X11) Tiene sentido analizar la relación entre el nivel de alcalinidad (X3) con el máximo de concentración de mercurio en cada grupo de peces (X10).

No tiene sentido analizar la concentración media de mercurio (X7) con el máximo de concentración de mercurio en cada grupo de peces (X10). No tiene sentido analizar la concentración media de mercurio (X7) con el mínimo de concentración de mercurio en cada grupo de peces (X9). No tiene sentido analizar el máximo de concentracón de mercurio (X10) con el mínimo de concentración de mercurio (X19). No tiene sentido analizar la relación entre la estimación de la concentración de mercurio en el pez de 3 años (X11) con la concentración media en el tejido muscular (X7). No tiene sentido analizar la relación entre la estimación de la concentración de mercurio en el pez de 3 años (X11) con el máximo de concentración de mercurio en cada grupo de peces (X10).

## 2. ANALIZA LOS DATOS Y PREGUNTA BASE

### 1. De acuerdo con la pregunta base, contempla la herramienta estadística necesaria para contestarla y justifica su elección.

La herramienta estadística necesaria para poder contestar la pregunta base es realizar una regresión lineal múltiple en donde se pueda analizar a mayor profundidad la relación lineal entre los factores propuestos sobre el nivel de contaminación de mercurio en los peces de los lagos de Florida para saber si realmente son significativas para determinar el nivel de contaminación de cada lago. De acuerdo a la matriz de correlación

elaborada en el inciso anterior, se determinó que las dos variables que tenían mayor relación con la estimación de mercurio en los peces fueron precisamente la alcalinidad (X3) y el pH (X4). Adicionalmente, se realizarán pruebas de hipótesis para analizar y evaluar si eventualmente, se está cumpliendo todas las condiciones para justificar la validez del modelo y reafirmar la relación encontrada entre las variables al presentar características similares a las de una distribución normal.

## 2. Aplicación de la herramienta estadística.

### 1. Análisis de Correlación de las variables

Matriz de Correlación y valores de probabilidad

```
#install.packages(psych)
library(psych)

## Warning: package 'psych' was built under R version 4.1.3

## Registered S3 method overwritten by 'psych':
##   method      from
##   plot.residuals rmutl

##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##   %+%, alpha

Rc = corr.test(df_num)
Rc
```

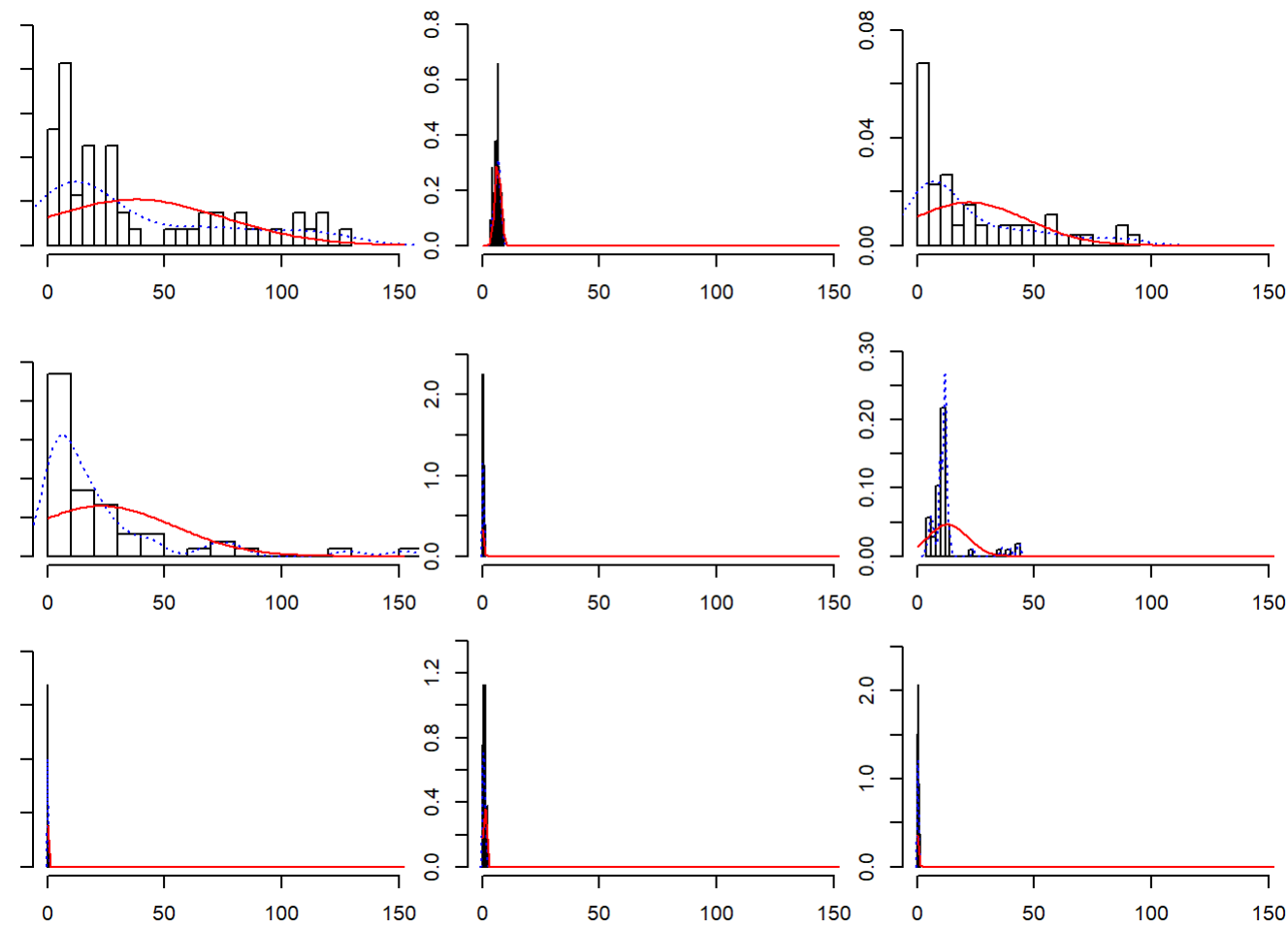


```
## Call:corr.test(x = df_num)
## Correlation matrix
##      X3    X4    X5    X6    X7    X8    X9    X10    X11
## X3  1.00  0.72  0.83  0.48 -0.59  0.01 -0.53 -0.60 -0.63
## X4  0.72  1.00  0.58  0.61 -0.58 -0.02 -0.54 -0.55 -0.61
## X5  0.83  0.58  1.00  0.41 -0.40 -0.09 -0.33 -0.41 -0.46
## X6  0.48  0.61  0.41  1.00 -0.49 -0.01 -0.40 -0.48 -0.51
## X7 -0.59 -0.58 -0.40 -0.49  1.00  0.08  0.93  0.92  0.96
## X8  0.01 -0.02 -0.09 -0.01  0.08  1.00 -0.08  0.16  0.03
## X9 -0.53 -0.54 -0.33 -0.40  0.93 -0.08  1.00  0.77  0.92
## X10 -0.60 -0.55 -0.41 -0.48  0.92  0.16  0.77  1.00  0.86
## X11 -0.63 -0.61 -0.46 -0.51  0.96  0.03  0.92  0.86  1.00
## Sample Size
## [1] 53
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##      X3    X4    X5    X6    X7    X8    X9    X10    X11
## X3  0.00  0.00  0.00  0.00  0.00  1.00  0.00  0.00  0.00
## X4  0.00  0.00  0.00  0.00  0.00  1.00  0.00  0.00  0.00
## X5  0.00  0.00  0.00  0.03  0.03  1.00  0.13  0.03  0.01
## X6  0.00  0.00  0.00  0.00  0.00  1.00  0.03  0.00  0.00
## X7  0.00  0.00  0.00  0.00  0.00  1.00  0.00  0.00  0.00
## X8  0.94  0.89  0.52  0.93  0.57  0.00  1.00  1.00  1.00
## X9  0.00  0.00  0.01  0.00  0.00  0.56  0.00  0.00  0.00
## X10 0.00  0.00  0.00  0.00  0.00  0.25  0.00  0.00  0.00
## X11 0.00  0.00  0.00  0.00  0.00  0.85  0.00  0.00  0.00
##
## To see confidence intervals of the correlations, print with the short=FALSE option
```

A partir de los valores-p entre las variables, se puede determinar que las más significativas con respecto a la concentración media de mercurio en los peces (X7) son: \* X3: Alcalinidad \* X4: pH \* X5: Calcio \* X6: Clorofila \* X9: Mínimo de Concentración de Mercurio en el grupo de peces \* X10: Máximo de Concentración de Mercurio en el grupo de peces \* X11: Estimación de la concentración de mercurio en el grupo de peces.

## Histogramas

```
library(psych)
multi.hist(x = df_num, dcol = c("blue", "red"), dlty = c("dotted", "solid"),
           main = "")
```



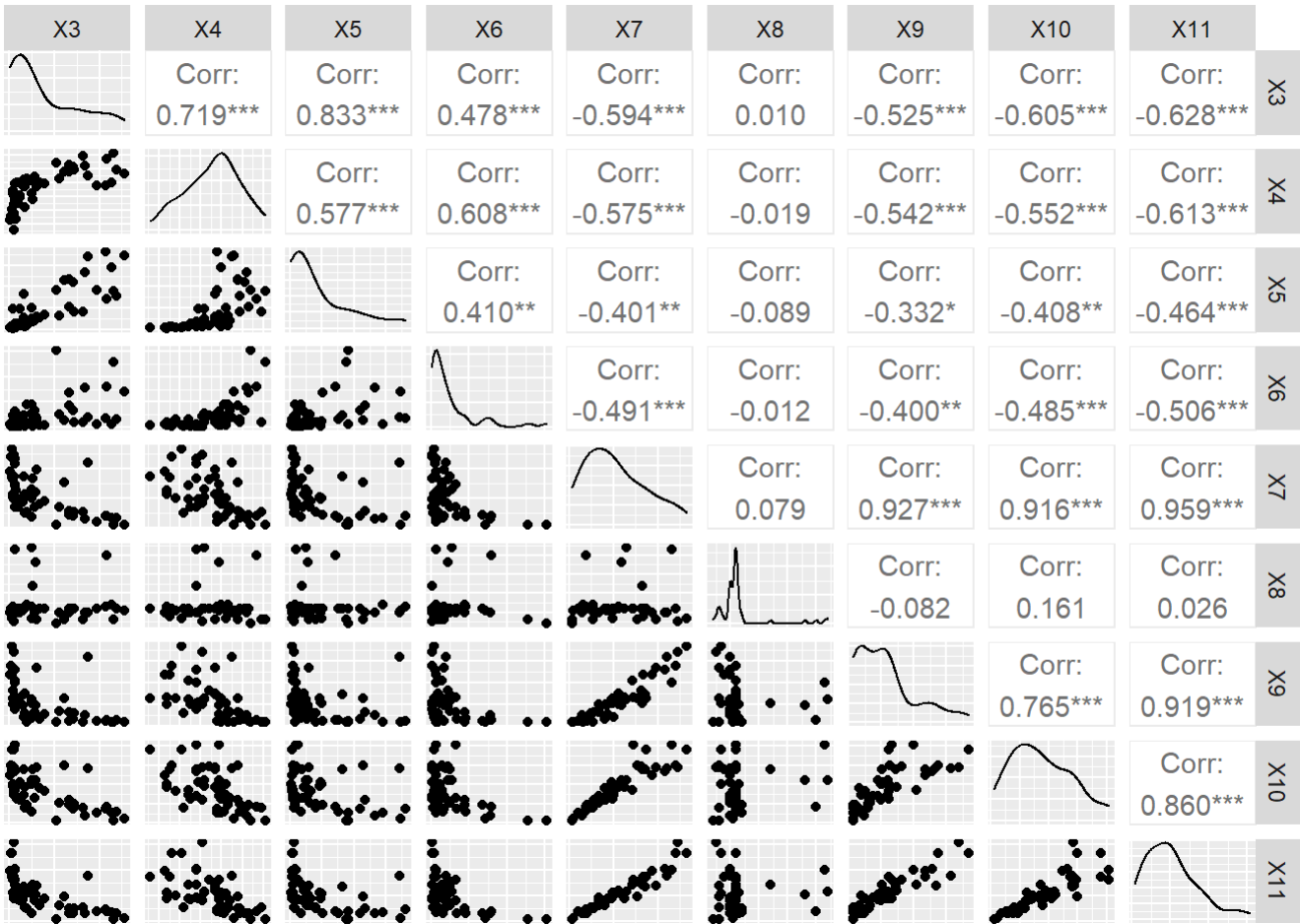
### Graficar visualmente las correlaciones

```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.1.3
```

```
## Registered S3 method overwritten by 'GGally':  
##   method from  
##   +.gg    ggplot2
```

```
ggpairs(df_num, lower=list(continuous="smooth"), diag = list(continuous="barDiag"), axisLabels="none")
```



A partir de la gráfica superior, se

puede ver de manera visual la dispersión de los datos en la relación de cada combinación de variables. Por lo que se puede ver una tendencia lineal positiva muy marcada entre X7 y X9 (Mínimo de concentración en los peces del lago), seguida de la tendencia de relación X7-X10 (Máximo de concentración en los peces del lago) y X7-X11 (Estimaciónj de concentración en los peces del lago). Al contrario, se observa gráficos donde claramente no existe una tendencia marcada entre las variables, como son las relaciones: X7-X8 (Número de peces en el lago), X7-X5 (Nivel de Calcio en el lago) y X7-X6 (Nivel de clorofila en el lago).

De igual manera, existen relaciones donde se aprecia una leve tendencia, no muy marcada entre los datos como son los casos de: X7-X3 (Alcalinidad en el lago) y X7-X4 (pH en el lago).

### Análisis del Modelo propuesto

```
modelo = lm(df_num$X7~df_num$X3+df_num$X4+df_num$X5+df_num$X6+df_num$X8+df_num$X9+df_num$X10+ df_num$X11, data=df_num)
summary(modelo)
```

```
##
## Call:
## lm(formula = df_num$X7 ~ df_num$X3 + df_num$X4 + df_num$X5 +
##     df_num$X6 + df_num$X8 + df_num$X9 + df_num$X10 + df_num$X11,
##     data = df_num)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.183976 -0.029092 -0.000945  0.026954  0.109588
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0728040  0.0701245  -1.038  0.304844
## df_num$X3    -0.0001311  0.0005194  -0.253  0.801826
## df_num$X4     0.0088597  0.0105876   0.837  0.407229
## df_num$X5     0.0003402  0.0006504   0.523  0.603527
## df_num$X6    -0.0003705  0.0003562  -1.040  0.303869
## df_num$X8     0.0019033  0.0010731   1.774  0.083062 .
## df_num$X9     0.5754473  0.1016125   5.663 1.05e-06 ***
## df_num$X10    0.2315867  0.0342125   6.769 2.49e-08 ***
## df_num$X11    0.3100547  0.0861305   3.600 0.000804 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05976 on 44 degrees of freedom
## Multiple R-squared:  0.974, Adjusted R-squared:  0.9693
## F-statistic: 206.2 on 8 and 44 DF, p-value: < 2.2e-16
```

El modelo con todas las variables introducidas como predictores tiene un R2 muy alta (0.974), lo que quiere decir que es capaz de explicar el 97,40% de la variabilidad observada en la contaminación de mercurio en los peces de los lagos de Florida. El p-value del modelo es significativo (2.2e-16) por lo que se puede aceptar que el modelo no se construyó aleatoriamente. Esto se debe a que, al menos uno de los coeficientes parciales de regresión es distinto de 0. Muchos de ellos no son significativos, lo que es un indicativo de que podrían no contribuir al modelo.

## Selección de los mejores predictores

Se utilizará la estrategia de stepwise mixto y el parámetro de Akaike(AIC) para evaluar la calidad del modelo.

```
step(object = modelo, direction = "both", trace = 1)
```

```

## Start:  AIC=-290.52
## df_num$X7 ~ df_num$X3 + df_num$X4 + df_num$X5 + df_num$X6 + df_num$X8 +
##      df_num$X9 + df_num$X10 + df_num$X11
##
##           Df Sum of Sq    RSS    AIC
## - df_num$X3   1  0.000228 0.15734 -292.44
## - df_num$X5   1  0.000977 0.15809 -292.19
## - df_num$X4   1  0.002500 0.15961 -291.68
## - df_num$X6   1  0.003865 0.16098 -291.23
## <none>                                0.15711 -290.52
## - df_num$X8   1  0.011231 0.16834 -288.86
## - df_num$X11  1  0.046272 0.20338 -278.84
## - df_num$X9   1  0.114518 0.27163 -263.50
## - df_num$X10  1  0.163612 0.32072 -254.70
##
## Step:  AIC=-292.44
## df_num$X7 ~ df_num$X4 + df_num$X5 + df_num$X6 + df_num$X8 + df_num$X9 +
##      df_num$X10 + df_num$X11
##
##           Df Sum of Sq    RSS    AIC
## - df_num$X5   1  0.000911 0.15825 -294.13
## - df_num$X4   1  0.002292 0.15963 -293.67
## - df_num$X6   1  0.003724 0.16106 -293.20
## <none>                                0.15734 -292.44
## - df_num$X8   1  0.011115 0.16845 -290.82
## + df_num$X3   1  0.000228 0.15711 -290.52
## - df_num$X11  1  0.046115 0.20345 -280.82
## - df_num$X9   1  0.114763 0.27210 -265.41
## - df_num$X10  1  0.184480 0.34182 -253.32
##
## Step:  AIC=-294.13
## df_num$X7 ~ df_num$X4 + df_num$X6 + df_num$X8 + df_num$X9 + df_num$X10 +
##      df_num$X11
##
##           Df Sum of Sq    RSS    AIC
## - df_num$X6   1  0.003671 0.16192 -294.92
## - df_num$X4   1  0.004107 0.16236 -294.78
## <none>                                0.15825 -294.13
## - df_num$X8   1  0.010893 0.16914 -292.61
## + df_num$X5   1  0.000911 0.15734 -292.44
## + df_num$X3   1  0.000162 0.15809 -292.19
## - df_num$X11  1  0.046058 0.20431 -282.60
## - df_num$X9   1  0.127701 0.28595 -264.78
## - df_num$X10  1  0.186249 0.34450 -254.91
##
## Step:  AIC=-294.92
## df_num$X7 ~ df_num$X4 + df_num$X8 + df_num$X9 + df_num$X10 +
##      df_num$X11
##
##           Df Sum of Sq    RSS    AIC
## - df_num$X4   1  0.001774 0.16370 -296.34
## <none>                                0.16192 -294.92
## + df_num$X6   1  0.003671 0.15825 -294.13
## - df_num$X8   1  0.010011 0.17193 -293.74
## + df_num$X5   1  0.000858 0.16106 -293.20

```

```
## + df_num$X3    1  0.000249 0.16167 -293.00
## - df_num$X11   1  0.052293 0.21422 -282.09
## - df_num$X9    1  0.124057 0.28598 -266.77
## - df_num$X10   1  0.192747 0.35467 -255.36
##
## Step:  AIC=-296.34
## df_num$X7 ~ df_num$X8 + df_num$X9 + df_num$X10 + df_num$X11
##
##           Df Sum of Sq    RSS    AIC
## <none>                 0.16370 -296.34
## - df_num$X8    1  0.010273 0.17397 -295.12
## + df_num$X5    1  0.001956 0.16174 -294.98
## + df_num$X4    1  0.001774 0.16192 -294.92
## + df_num$X6    1  0.001338 0.16236 -294.78
## + df_num$X3    1  0.001299 0.16240 -294.76
## - df_num$X11   1  0.050769 0.21446 -284.02
## - df_num$X9    1  0.126611 0.29031 -267.98
## - df_num$X10   1  0.191228 0.35492 -257.33
```

```
##
## Call:
## lm(formula = df_num$X7 ~ df_num$X8 + df_num$X9 + df_num$X10 +
##     df_num$X11, data = df_num)
##
## Coefficients:
## (Intercept)    df_num$X8    df_num$X9    df_num$X10    df_num$X11
##    -0.01651      0.00176      0.57486      0.23673      0.29777
```

Por ende, el mejor modelo y conjunto de predictores resultante del proceso ha sido:

```
modelo = lm(formula =df_num$X7 ~ df_num$X8 + df_num$X9 + df_num$X10 + df_num$X11, data = df_num)
summary(modelo)
```

```
##
## Call:
## lm(formula = df_num$X7 ~ df_num$X8 + df_num$X9 + df_num$X10 +
##     df_num$X11, data = df_num)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.188438 -0.028797 -0.003159  0.030784  0.117510
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.016513   0.019166  -0.862   0.39321
## df_num$X8    0.001760   0.001014   1.736   0.08906 .
## df_num$X9    0.574864   0.094347   6.093 1.81e-07 ***
## df_num$X10   0.236728   0.031613   7.488 1.32e-09 ***
## df_num$X11   0.297773   0.077177   3.858 0.00034 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0584 on 48 degrees of freedom
## Multiple R-squared:  0.9729, Adjusted R-squared:  0.9707
## F-statistic: 431.3 on 4 and 48 DF,  p-value: < 2.2e-16
```

Con las variables predictoras: \* X8: Cantidad de peces en el lago \* X9: mínimo de la concentración de mercurio en cada grupo de peces \* X10: Máximo de la concentración de mercurio en cada grupo de peces \* X11: Estimación de la concentración de mercurio en cada grupo de peces

Intervalos de confianza para cada coeficiente parcial de regresión

```
confint(modelo)

##              2.5 %       97.5 %
## (Intercept) -0.0550486624 0.022023032
## df_num$X8    -0.0002789989 0.003799813
## df_num$X9     0.3851664286 0.764561392
## df_num$X10    0.1731645931 0.300290751
## df_num$X11    0.1425985956 0.452946867
```

Cada una de las pendientes de un modelo de regresión lineal múltiple (mayormnente conocidos como coeficientes parciales de regresión de los predictores) se define del siguiente modo: Si el resto de variables se mantienen constantes, por cada unidad que aumenta el predictor en cuestión, la variable (Y) varía en promedio tantas unidades como indica la pendiente.

Por cada unidad que aumenta el predictor X8 (Número de peces en el lago), la concentración media de mercurio en los peces del lago aumenta en promedio 0.00176041 unidades, manteniéndose constantes el resto de predictores.

Por cada unidad que aumenta el predictor X9 (Concentración mínima de mercurio en el lago), la concentración media de mercurio en los peces del lago aumenta en promedio 0.5749 unidades, manteniéndose constantes el resto de predictores.

Por cada unidad que aumenta el predictor X10 (Concentración máxima de mercurio en el lago), la concentración media de mercurio en los peces del lago aumenta en promedio 0.2367 unidades, manteniéndose constantes el resto de predictores.

Por cada unidad que aumenta el predictor X11 (Estimación de la concentración promedio de mercurio en el lago), la concentración media de mercurio en los peces del lago aumenta en promedio 0.2977 unidades, manteniéndose constantes el resto de predictores.

### 3. Valida el modelo obtenido analizando los supuestos requeridos por el modelo.

### Verifica que  $\beta_1$  sea significativa con un alfa de 0.05.  $H_0 = \beta_1$  no es significativa.  $H_1 = \beta_1$  es significativa.

Obtener el summary de ambas correlaciones

```
#Para obtener t estrella
summary(modelo)
```

```
##
## Call:
## lm(formula = df_num$X7 ~ df_num$X8 + df_num$X9 + df_num$X10 +
##     df_num$X11, data = df_num)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.188438 -0.028797 -0.003159  0.030784  0.117510
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.016513   0.019166  -0.862   0.39321
## df_num$X8     0.001760   0.001014   1.736   0.08906 .
## df_num$X9     0.574864   0.094347   6.093 1.81e-07 ***
## df_num$X10    0.236728   0.031613   7.488 1.32e-09 ***
## df_num$X11    0.297773   0.077177   3.858 0.00034 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0584 on 48 degrees of freedom
## Multiple R-squared:  0.9729, Adjusted R-squared:  0.9707
## F-statistic: 431.3 on 4 and 48 DF,  p-value: < 2.2e-16
```

```
te_X8 = 1.736
te_X9 = 6.093
te_X10 = 7.488
te_X11 = 3.858

pvalue_H = 2.2e-16
```

Obtener  $t_0$  con tstudent

```
alpha = 0.05
t_0 = qt(alpha/2, n-2)
print(t_0)
```

```
## [1] -2.007584
```

Interpretación en el contexto del problema:

Se rechaza la hipótesis nula  $H_0$  ya que: En primera instancia, los valores de  $t_{\text{estrella}}$  de cada predictor son mayores al valor de  $t_0$ . En segunda instancia, los valores-p obtenidos son menores a  $\alpha$  (0.05)



Por lo tanto, se puede llegar a la conclusión de que los valores de  $\beta_i$  si son significativos, es decir que todos tienen un valor distinto de 0 y que, por lo tanto, sí son relevantes para el análisis y cómo supera el valor de alpha, se extiende de la región de aceptación de la hipótesis nula, por lo tanto, se deberá rechazarla.

Dentro del contexto del problema, los valores de  $\beta_i$  son la tasa de cambio a la que varía la concentración de Mercurio en los peces del lago en función de cada una de las variables predictoras, respectivamente. Por ende, se sabe que sí existe una relación de dependencia lineal entre las variables y que dicha relación es constante y diferente de 0 ya que la recta es la de mejor ajuste porque supera el nivel de significancia de alpha. Entonces, dicho modelo representa correctamente la relación entre las variables. No necesariamente indica causalidad, pero sí relación entre las variables.

## No colinealidad de xi

Relación lineal entre los predictores numéricos y la variable respuesta:

```
library(ggplot2)
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 4.1.3
```

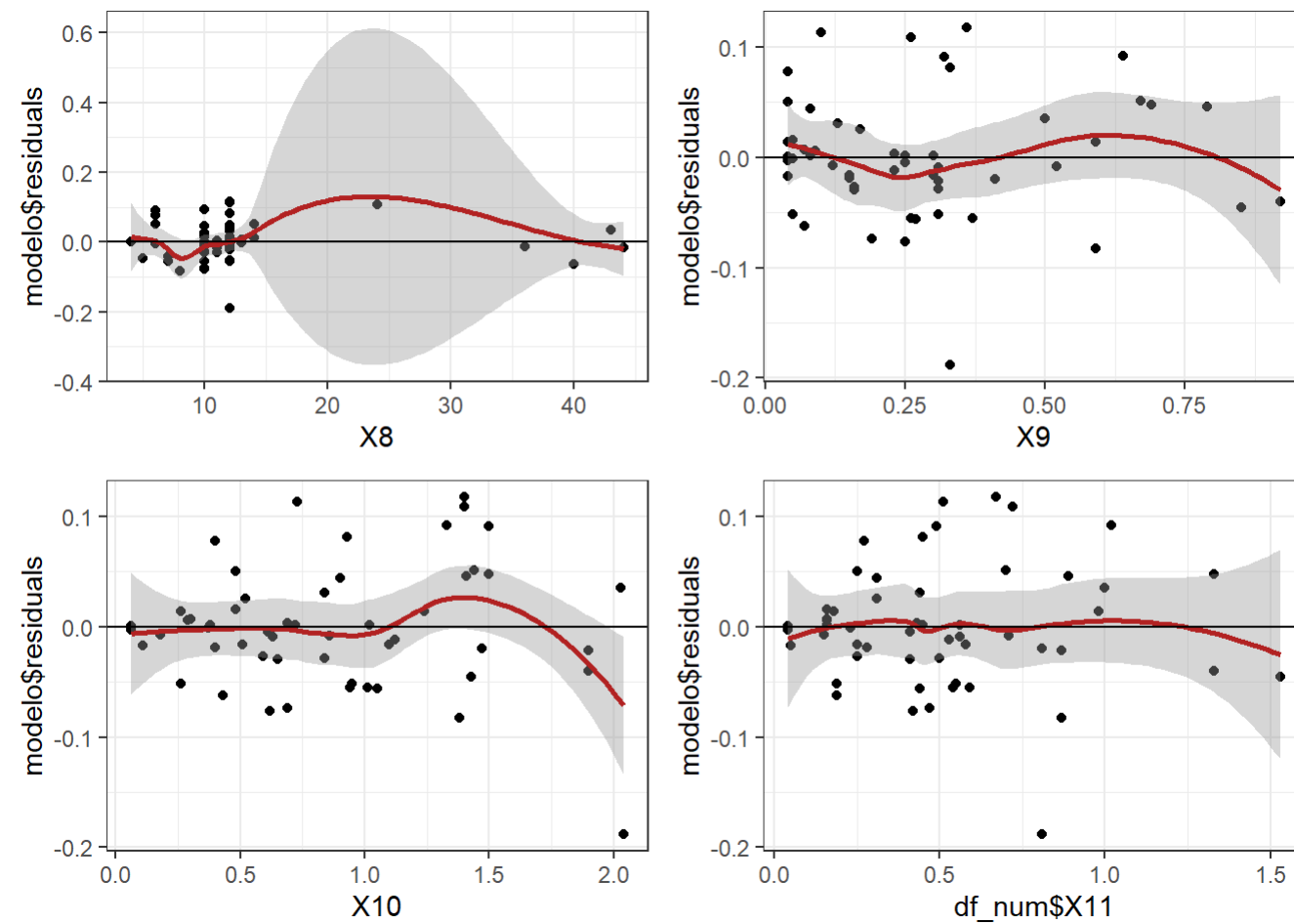
```
plot1 <- ggplot(data = df_num, aes(X8, modelo$residuals)) +
  geom_point() + geom_smooth(color = "firebrick") + geom_hline(yintercept = 0) +
  theme_bw()
plot2 <- ggplot(data = df_num, aes(X9, modelo$residuals)) +
  geom_point() + geom_smooth(color = "firebrick") + geom_hline(yintercept = 0) +
  theme_bw()
plot3 <- ggplot(data = df_num, aes(X10, modelo$residuals)) +
  geom_point() + geom_smooth(color = "firebrick") + geom_hline(yintercept = 0) +
  theme_bw()
plot4 <- ggplot(data = df_num, aes(df_num$X11, modelo$residuals)) +
  geom_point() + geom_smooth(color = "firebrick") + geom_hline(yintercept = 0) +
  theme_bw()
grid.arrange(plot1, plot2, plot3, plot4)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## Warning: Use of `df_num$X11` is discouraged. Use `X11` instead.
```

```
## Warning: Use of `df_num$X11` is discouraged. Use `X11` instead.
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

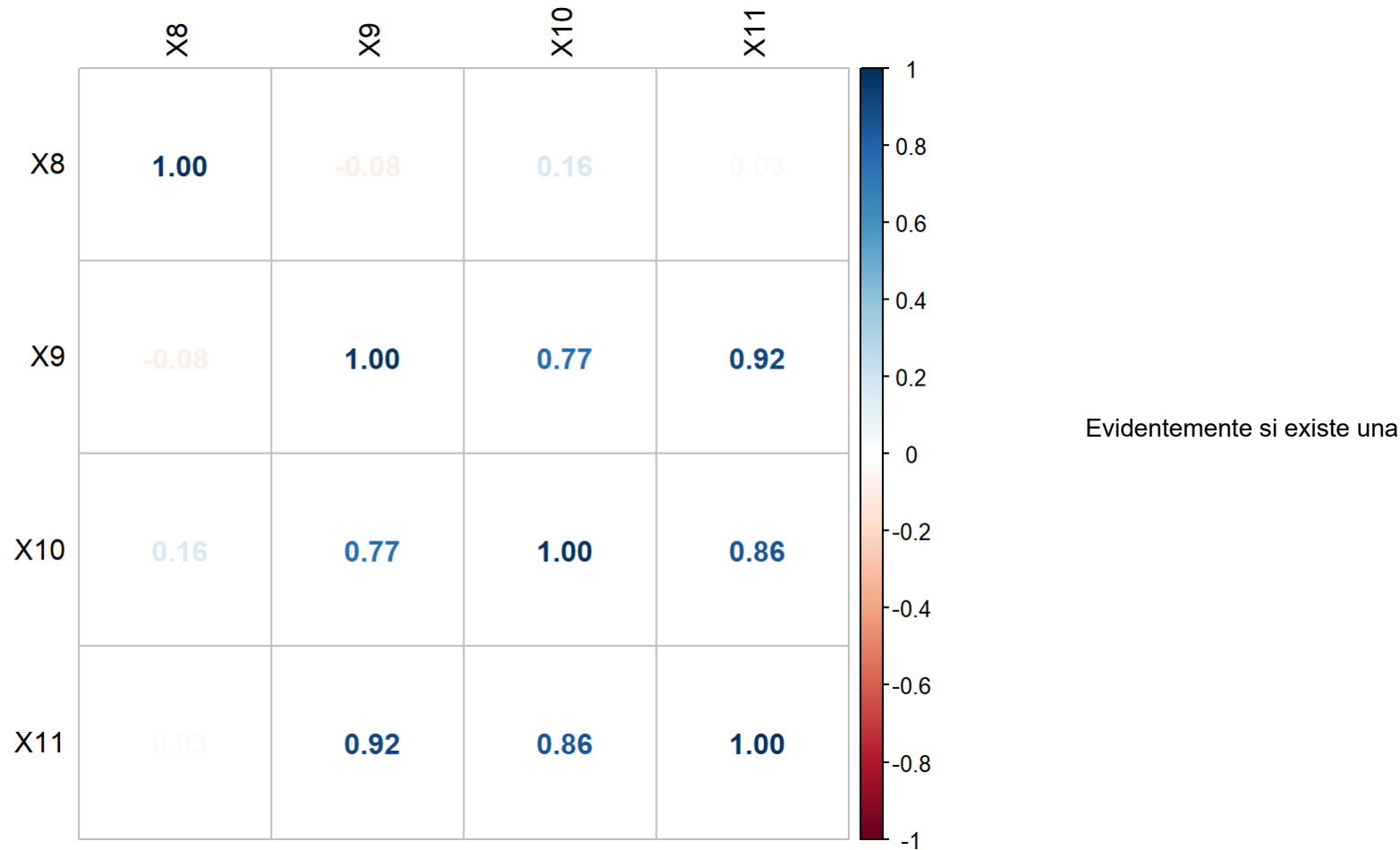


En esta gráfica se está observando

la dispersión entre cada uno de los predictores con los residuos del modelo planteado. Para cumplir con este primer supuesto, los residuos deberán distribuirse de manera aleatoria alrededor de 0 y demostrando una variabilidad constante a lo largo del eje X. Por lo que, solamente 3 de los 4 predictores cumplen esta condición, mientras que el gráfico de X8 (Cantidad de peces en el lago) no tiene una variabilidad constante a lo largo del eje horizontal, sino que está muy concentrada en el primer tramo. Por lo que no se está presentando linealidad en todos los predictores.

## No multicolinealidad:

```
library(corrplot)
corrplot(cor(dplyr::select(df_num, X8, X9, X10, X11)),
          method = "number", tl.col = "black")
```



Evidentemente si existe una correlación alta entre las variables predictoras X9, X10 y X11. Dentro del contexto del problema, esto significa que existe una relación lineal de dependencia entre el mínimo de concentración de mercurio en el grupo de peces del lago, el máximo de concentración de mercurio en el grupo de peces del lago y la estimación de concentración media de mercurio en el lago, lo cuál suena bastante lógico ya que estas 3 variables hacen referencia a la misma métrica base (concentración de mercurio). No obstante, para términos de ser consideradas como factores determinantes del nivel de contaminación en los lagos, no son las más adecuadas.

Análisis de Inflación de Varianza (VIF):

```
library(car)

## Warning: package 'car' was built under R version 4.1.3

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.1.3

##
## Attaching package: 'car'

## The following object is masked from 'package:psych':
##
##   logit

vif(modelo)
```

```
## df_num$X8 df_num$X9 df_num$X10 df_num$X11
## 1.149643 6.957250 4.153061 10.420386
```

Existe un nivel de varianza considerable en las variables predictoras lo que indica una intensidad media de multicolinealidad. Dentro del contexto del problema, esto significa que dichos coeficientes se incrementan en un promedio de 5 puntos a causa de la colinealidad. La única variable que presenta un nivel de varianza alto (>10) es la variable X11 (estimación de la concentración de mercurio promedio), por lo cuál se infiere que esta es la variable que más se ve afectada por la colinealidad.

### Autocorrelación:

```
library(car)
dwt(modelo, alternative = "two.sided")
```

```
## lag Autocorrelation D-W Statistic p-value
## 1 0.1411273 1.703788 0.262
## Alternative hypothesis: rho != 0
```

No hay evidencia de autocorrelación.

### Tamaño de la muestra:

```
print(paste("La cantidad de datos es de: ", n))
```

```
## [1] "La cantidad de datos es de: 53"
```

```
print("La cantidad de predictores es de: 4")
```

```
## [1] "La cantidad de predictores es de: 4"
```

Aunque no exista un criterio establecido para el número mínimo de observaciones, para prevenir que una variable resulte muy influyente cuando realmente no lo es, se recomienda que la cantidad de observaciones sea entre 10 y 20 veces el número de predictores. En este caso debería haber como mínimo 40 observaciones y se dispone de 53 por lo que es apropiado.

### ##Economía de las variables: Coeficiente de Determinación

```
summary(modelo)
```

```
##
## Call:
## lm(formula = df_num$X7 ~ df_num$X8 + df_num$X9 + df_num$X10 +
##     df_num$X11, data = df_num)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.188438 -0.028797 -0.003159  0.030784  0.117510
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.016513   0.019166  -0.862   0.39321
## df_num$X8     0.001760   0.001014   1.736   0.08906 .
## df_num$X9     0.574864   0.094347   6.093 1.81e-07 ***
## df_num$X10    0.236728   0.031613   7.488 1.32e-09 ***
## df_num$X11    0.297773   0.077177   3.858 0.00034 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0584 on 48 degrees of freedom
## Multiple R-squared:  0.9729, Adjusted R-squared:  0.9707
## F-statistic: 431.3 on 4 and 48 DF,  p-value: < 2.2e-16
```

Analizando el valor del coeficiente de determinación, se puede ver que el modelo se ajusta casi perfectamente a las observaciones reales ya que su valor oscila dentro del rango de correlaciones extremadamente fuertes, por lo que se puede considerar como un excelente modelo para describir los datos. Además, se realizó un proceso previo de selección de modelo aplicando la estrategia fde stepwise mixto con el parámetro de Akaike para seleccionar las variables predictoras más influyentes que tuvieran el nivel de significancia más elevado, por lo que se puede justificar el alto nivel del coeficiente de determinación.

###Análisis de los residuos

## Normalidad de los residuos

### 1) Hipótesis

H0 = Los datos provienen de una población normal. H1 = Los datos no provienen de una población normal.

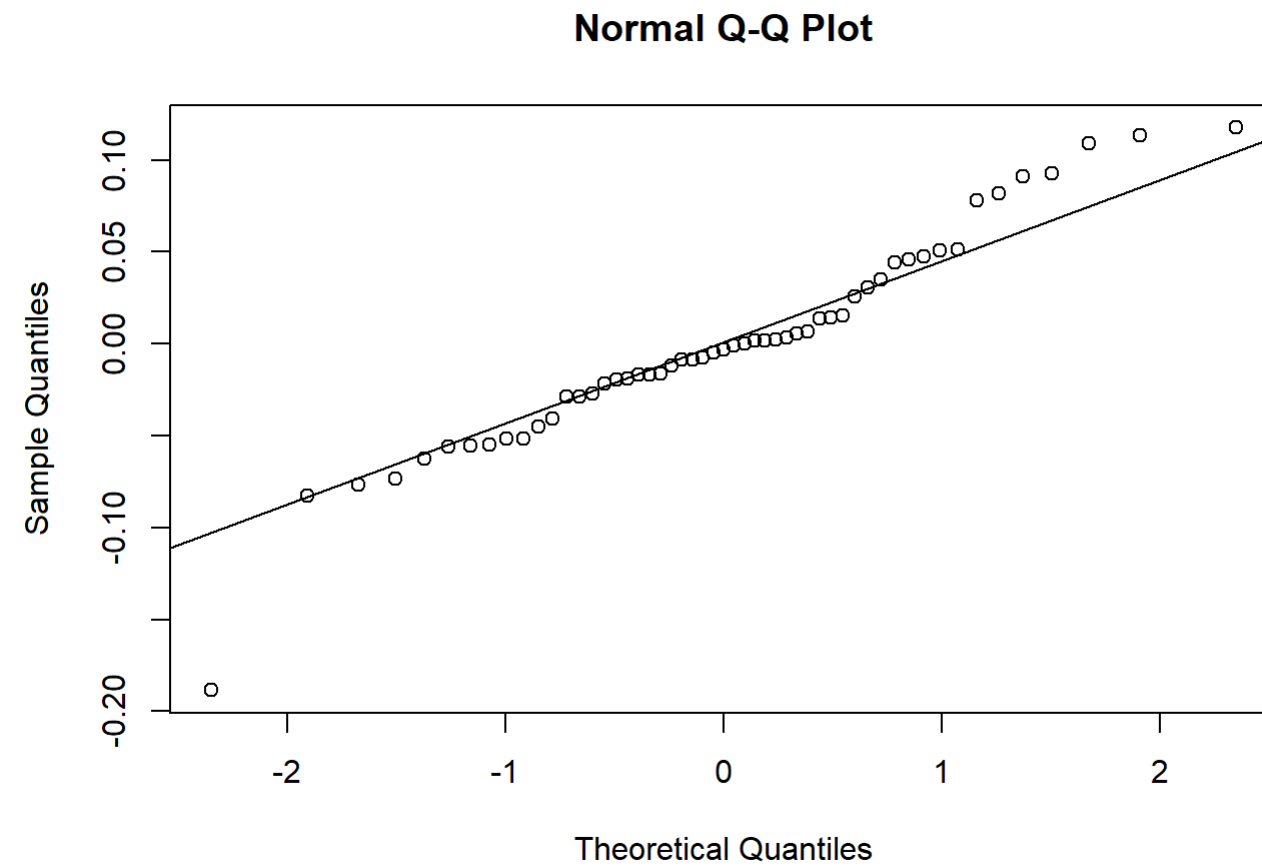
### 2) Regla de Desición

Se rechaza H0 si valor  $p < \alpha$

### 3) Análisis del Resultado

Distribución normal de los residuos:

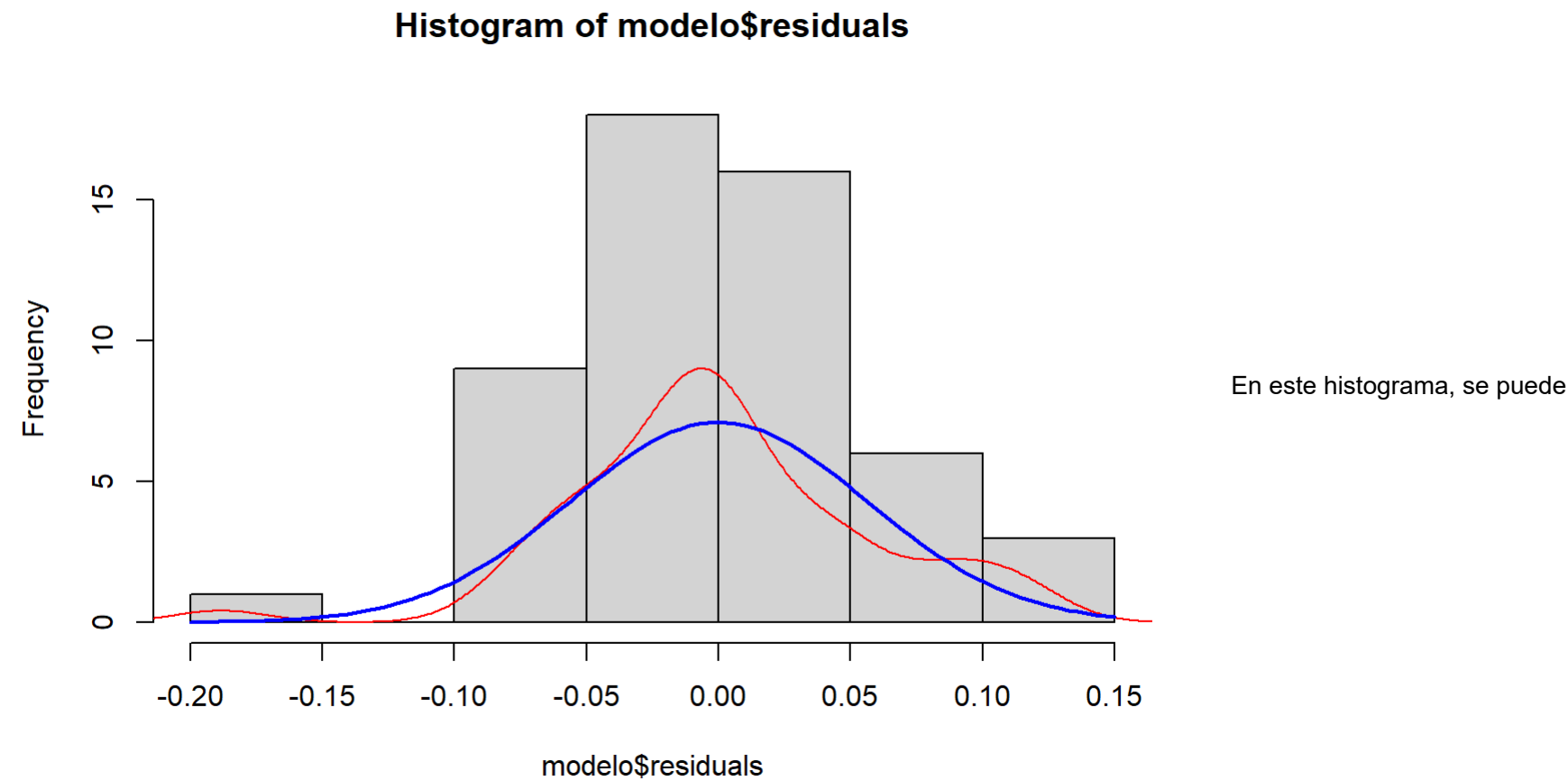
```
qqnorm(modelo$residuals)
qqline(modelo$residuals)
```



El comportamiento de esta gráfica de

probabilidad normal presenta una distribución con colas suaves la cual posee una alta curtosis en forma de una distribución Leptocúrtica. Los residuos en efecto, se comportan como una distribución normal ya que se ajustan casi perfectamente a una línea recta y tienen una tendencia creciente. De igual manera, se procedió a graficar un histograma de frecuencias para observar la distribución de la data.

```
hist(modelo$residuals)
lines(density(modelo$residual),col="red")
curve(dnorm(x,mean=mean(modelo$residuals)
,sd=sd(modelo$residuals)), add=TRUE, col="blue",lwd=2)
```



verificar claramente que la mayor agrupación de los datos está en el centro de la distribución y la menor proporción de los datos se encuentra en los extremos, por ende, se asemeja perfectamente a una distribución normal.

```
shapiro.test(modelo$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  modelo$residuals
## W = 0.95618, p-value = 0.0499
```

#### 4) Conclusión

El valor p de los residuos es igual al valor de alpha establecido (0.05), por lo que, se procede a aceptar la hipótesis nula  $H_0$ . Las gráficas de qq-plot y el histograma de distribución encontraron que los datos tienen simetría con un leve sesgo a la derecha y que la mayor parte de los datos está concentrada en la mitad y se ajustan perfectamente al modelo propuesto. De hecho, el histograma muestra claramente que la densidad de datos sigue una perfecta campana simétrica donde la mayor frecuencia de datos se ubica en torno a la media y la menor distribución se encuentra en los extremos. De modo que, se indica que los residuos siguen una distribución normal.

#### Verificación de media cero

##### 1) Hipótesis

$h_0$  = la media es 0.  $h_1$  = la media es diferente de 0.

##### 2) Reglas de Decisión

Se rechaza  $h_0$  si: Regla clásica: Si  $|t^*|$  es mayor a  $|t_0|$  Regla valor p: Si valor  $p < \alpha$

##### 3) Análisis de Resultado

```
t.test(modelo$residuals, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  modelo$residuals
## t = 4.1225e-17, df = 52, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.01546503  0.01546503
## sample estimates:
##    mean of x
## 3.177173e-19
```

#### 4) Conclusión

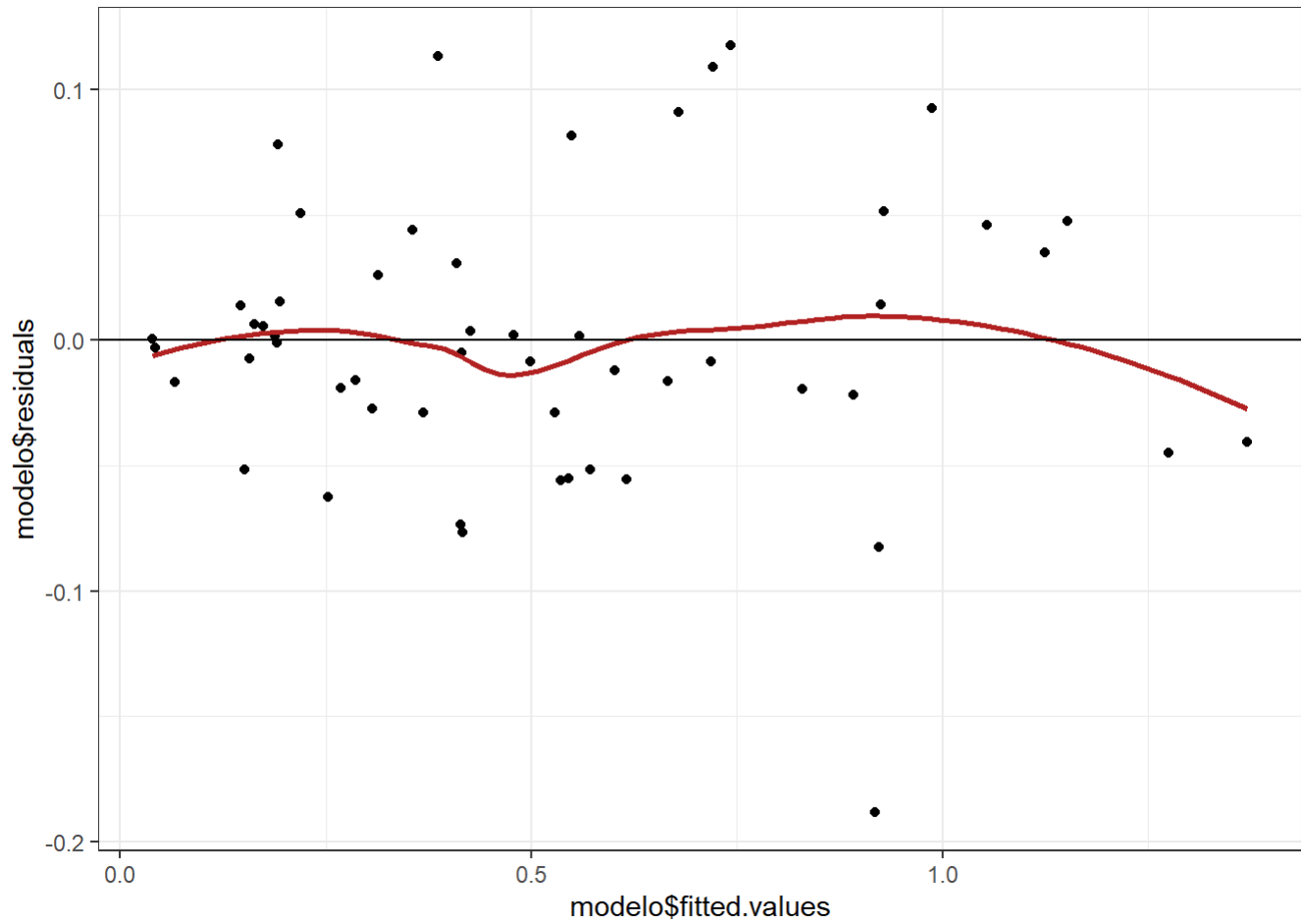
A través de esta prueba se demuestra que los valores de la media es diferente de 0. Por consiguiente, se procede a rechazar la hipótesis nula. Adicionalmente, se comprueba que los valores de  $|t^*|$  son mayores que el valor de  $|t_0|$  y que el valor-p (1) es mayor al valor propuesto de alpha (0.05). Entonces, se procede a rechazar la hipótesis nula. Por lo que la media del modelo no es igual a 0.

### Variabilidad constante de los residuos (homocedasticidad):

```
ggplot(data = df_num, aes(modelo$fitted.values, modelo$residuals)) +geom_point() +
geom_smooth(color = "firebrick", se = FALSE) +
geom_hline(yintercept = 0) +
theme_bw()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```





```
library(lmtest)

## Warning: package 'lmtest' was built under R version 4.1.3

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 4.1.3

##
## Attaching package: 'zoo'

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

bptest(modelo)

##
## studentized Breusch-Pagan test
##
## data:  modelo
## BP = 18.379, df = 4, p-value = 0.001041
```

A partir de este test, se demuestra que no hay evidencia de falta de homocedasticidad.

## Análisis del resultado

Una vez más, en la gráfica superior se puede apreciar claramente que los residuos se distribuyen de manera aleatoria en torno a 0 y que tienen aproximadamente una variabilidad constante a lo largo del eje horizontal ya que no tienen un patrón específico y que hay ausencia de dispersión en los extremos. Por ende, se puede constatar que la variabilidad es independiente del valor ajustado. En efecto, se puede denotar que no existe ninguna estructura evidente en los datos, por lo que la varianza de los residuos debe ser constante en casi sus valores. Aparte, se infiere claramente que hay una independencia entre las variables, entonces, se puede afirmar que hay linealidad y existencia de homocedasticidad. De modo que los residuos son aleatorios lo que implica que el modelo es adecuado y se reivindica su homocedasticidad.

## Análisis de datos atípicos o influyentes

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.3
```

```
##  
## Attaching package: 'dplyr'
```

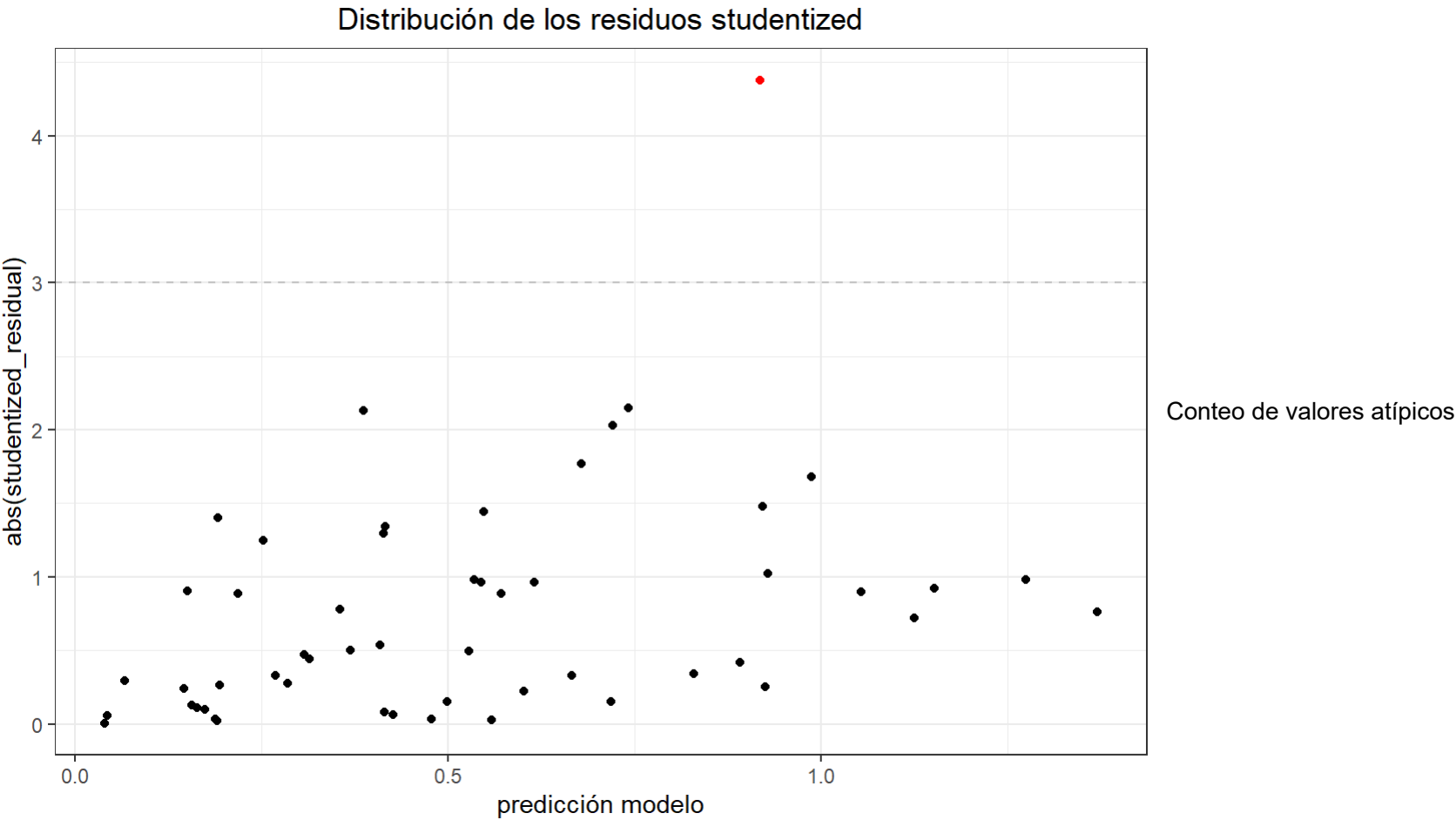
```
## The following object is masked from 'package:car':  
##  
##      recode
```

```
## The following object is masked from 'package:gridExtra':  
##  
##      combine
```

```
## The following objects are masked from 'package:stats':  
##  
##      filter, lag
```

```
## The following objects are masked from 'package:base':  
##  
##      intersect, setdiff, setequal, union
```

```
library(ggplot2)  
df_num$studentized_residual <- rstudent(modelo)  
ggplot(data = df_num, aes(x = predict(modelo), y = abs(studentized_residual))) +  
  geom_hline(yintercept = 3, color = "grey", linetype = "dashed") +  
  # se identifican en rojo observaciones con residuos estandarizados absolutos > 3  
  geom_point(aes(color = ifelse(abs(studentized_residual) > 3, 'red', 'black'))) +  
  scale_color_identity() +  
  labs(title = "Distribución de los residuos studentized",  
       x = "predicción modelo") +  
  theme_bw() + theme(plot.title = element_text(hjust = 0.5))
```



```
which(abs(df_num$studentized_residual) > 3)
```

```
## [1] 24
```

Se observan un valor atípico en el reglón 24 en el modelo que podría causando ruido en el modelo.

```
summary(influence.measures(modelo))
```

```
## Potentially influential observations of
## lm(formula = df_num$X7 ~ df_num$X8 + df_num$X9 + df_num$X10 + df_num$X11, data = df_num) :
##
##      dfb.1_ dfb.d_$X8 dfb.d_$X9 dfb.d_$X10 dfb.d_$X11 dffit  cov.r  cook.d
## 1  0.08  0.08  0.19  0.50 -0.57 -0.78  1.63_* 0.12
## 14 -0.34 0.36  0.01  0.12 -0.02  0.48  1.52_* 0.05
## 20 0.06 -0.07  0.50  0.27 -0.50  0.61  1.34_* 0.07
## 24 0.34 0.90  0.99 -2.08_* 0.19 -2.54_* 0.27_* 0.93_*
## 33 0.03 0.08  0.14 -0.14 -0.05 -0.22  1.40_* 0.01
## 40 -0.02 0.05  0.45  0.06 -0.32  0.50  1.34_* 0.05
## 47 0.10 -0.20 -0.04  0.02  0.02 -0.21  1.54_* 0.01
##      hat
## 1  0.38_*
## 14 0.31_*
## 20 0.26
## 24 0.25
## 33 0.22
## 40 0.24
## 47 0.29_*
```

## 2) Criterios de Decisión

En la tabla generada se recogen las observaciones que son significativamente influyentes en al menos uno de los predictores (una columna para cada predictor). Las tres últimas columnas son 3 medidas distintas para cuantificar la influencia. A modo de guía se pueden considerar excesivamente influyentes aquellas observaciones para las que:

Leverages (hat): Se consideran observaciones influyentes aquellas cuyos valores hat superen  $2.5((p+1)/n)$ , siendo p el número de predictores y n el número de observaciones. Distancia Cook (cook.d): Se consideran influyentes valores superiores a 1.

## 3) Análisis de Resultados

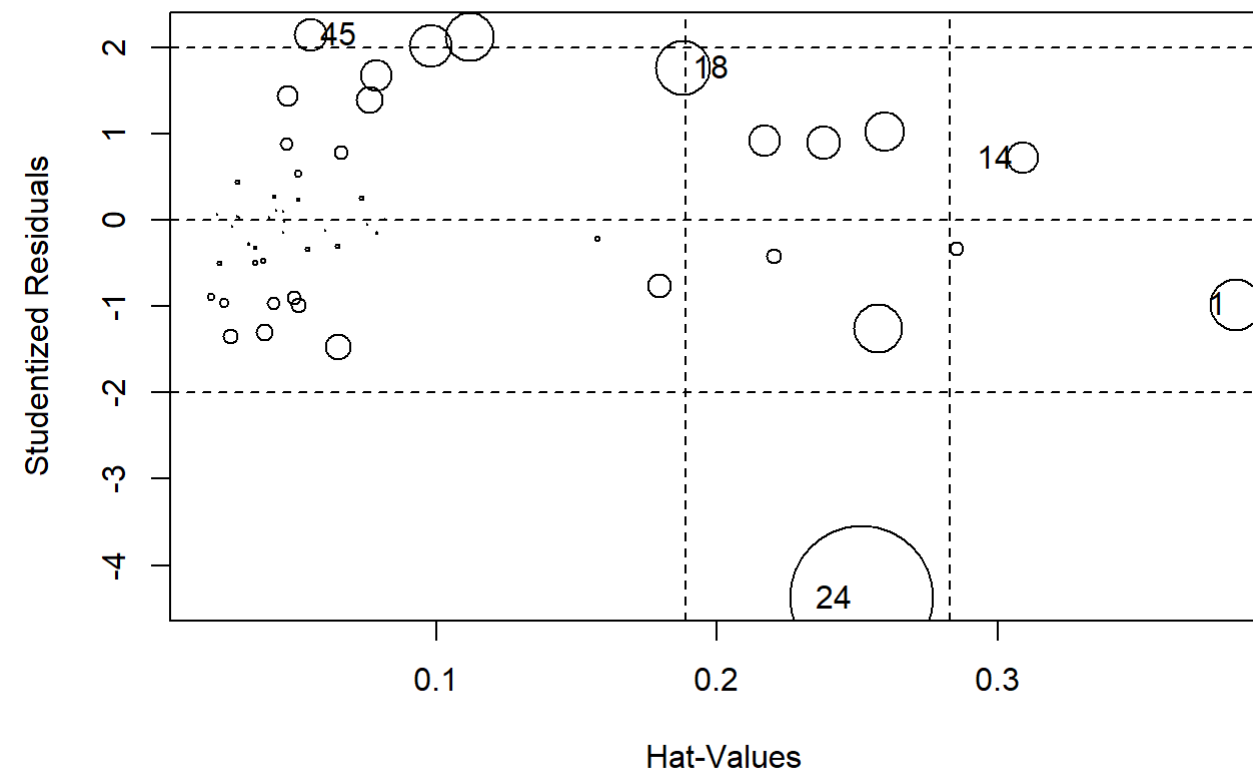
```
pred =4
hat_Value = 2.5*((pred+1)/n)
print(paste("El valor hat de referencia del modelo es: ", hat_Value))
```

```
## [1] "El valor hat de referencia del modelo es: 0.235849056603774"
```

De acuerdo al criterio de leverage, todas las observaciones marcadas en la tabla superior se consideran significativamente influyentes para el modelo. Mientras que en el criterio de Distancia cook, ninguna de ellas es lo suficientemente influyente como para predecir

####visualización gráfica de las influencias

```
influencePlot(modelo)
```



##	StudRes	Hat	CookD
## 1	-0.9828460	0.3847511	0.12090320
## 14	0.7183876	0.3088920	0.04660248
## 18	1.7666459	0.1877203	0.13815155
## 24	-4.3792946	0.2514812	0.93468910
## 45	2.1465084	0.0551522	0.05002932

El modelo lineal múltiple Concentración media de mercurio en los peces de los lagos de Florida = 0.001760X8 + 0.574864X9 + 0.2367X10 + 0.2978X11 es capaz de explicar el 97.29% de la variabilidad observada en la concentración media de mercurio (R2: 0.9729, R2-Adjusted: 0.9707). El test F muestra que es significativo (p-value: 2.2e-16). No se satisfacen todas las condiciones para este tipo de regresión múltiple ya que: \* La media de los residuos es diferente de 0. \* Existe un dato atípico que podría estar afectando la calidad del modelo. \* Existe correlación entre 3 variables predictoras (X9, X10, X11), lo cuál invalida el supuesto de no-colinealidad entre las variables independientes. \* La variable predictora X8 no presenta linealidad. \* La variable X11 tiene un alto grado de varianza por lo que está afecta en gran parte por la colinealidad.

## 7. Conclusión

¿Cuáles son los principales factores que influyen en el nivel de contaminación por mercurio en los peces de los lagos de Florida? De acuerdo al modelo propuesto, los factores principales que más influyen en el nivel de contaminación por mercurio en los peces de los lagos de Florida son: Cantidad de peces en el lago, Mínimo de concentración de mercurio en el grupo de peces, máximo de concentración de mercurio en el grupo de peces, y Estimación de la concentración de mercurio. No obstante, se recomienda ajustar el modelo para refinar los resultados y acotar variables que están produciendo colinealidad y/o redundancia para dar mayor confiabilidad y validez al modelo, además de explorar el efecto que tienen las demás variables que no son directamente dependientes de la concentración media de mercurio.