Bitacora 2

Cristhian, Olman, Jose

Integrantes

-Cristhian Jimenez Campos C33973

-Olman Camacho Jerez C31523

-Jose Manuel Alfaro Monge C30244

1- Identifique los siguientes puntos para los datos con los que realizará su proyecto:

Datos

Descripción de la tabla de datos.

Caracteristicas de la tabla

Esta base de datos contiene registros de 4424 estudiantes, quienes serán clasificados de distintas maneras, desde estado civil hasta los cursos que están cursando, entre otros. La base de datos fue publicada en 2021 y presenta diversos factores, incluyendo variables relacionadas con los padres, para analizar si influyen en la vida académica del estudiante. Fue creada por Valentim Realinho, Mónica Vieira Martins, Jorge Machado y Luís Baptista, investigadores del Instituto Politécnico de Portalegre en Portugal, y descargada desde el enlace link. Los datos corresponden al segundo semestre, aunque no se especifica el año.

Las variables están distribuidas en distintas categorías: variables relacionadas con la trayectoria académica, variables demográficas y variables socioeconómicas. Los tipos de datos incluyen variables reales, categóricas y enteras.

Poblacion de estudio:

Estudiantes matriculados en diferentes carreras de pregrado de una institucion de educacion superior.

Muestra observada.
4,424 estudiantes.

Unidad estadística o individuos.
Cada uno de los 4,424 estudiantes de educación superior durante determinados semestres.

Identificación de las variables de estudio.
Las variables de estudio incluyen información sobre la trayectoria académica, datos demográficos y factores socioeconómicos de los estudiantes, así como su rendimiento académico al final del primer y segundo semestre. El problema se plantea como una tarea de clasificación en tres categorías: abandono, matriculado y graduado.

2- primeas 5 filas de la tabla de datos

```
library(dplyr)
```

```
Adjuntando el paquete: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
library(ggplot2)
datos <- read.csv2("data.csv", sep = ";", header = TRUE, stringsAsFactors = FALSE)
head(datos, 5)
```

```
  Marital.status Application.mode Application.order Course
1              1               17                 5    171
2              1               15                 1   9254
3              1                1                 5   9070
4              1               17                 2   9773
5              2               39                 1   8014
  Daytime.evening.attendance. Previous.qualification
1                           1                      1
2                           1                      1
3                           1                      1
4                           1                      1
5                           0                      1
```

```
  Previous.qualification..grade. Nacionality Mother.s.qualification
1                          122.0           1                      19
2                          160.0           1                       1
3                          122.0           1                      37
4                          122.0           1                      38
5                          100.0           1                      37
  Father.s.qualification Mother.s.occupation Father.s.occupation
1                     12                   5                   9
2                      3                   3                   3
3                     37                   9                   9
4                     37                   5                   3
5                     38                   9                   9
  Admission.grade Displaced Educational.special.needs Debtor
1           127.3         1                         0      0
2           142.5         1                         0      0
3           124.8         1                         0      0
4           119.6         1                         0      0
5           141.5         0                         0      0
  Tuition.fees.up.to.date Gender Scholarship.holder Age.at.enrollment
1                       1      1                  0                20
2                       0      1                  0                19
3                       0      1                  0                19
4                       1      0                  0                20
5                       1      0                  0                45
  International Curricular.units.1st.sem..credited.
1             0                                   0
2             0                                   0
3             0                                   0
4             0                                   0
5             0                                   0
  Curricular.units.1st.sem..enrolled. Curricular.units.1st.sem..evaluations.
1                                   0                                      0
2                                   6                                      6
3                                   6                                      0
4                                   6                                      8
5                                   6                                      9
  Curricular.units.1st.sem..approved. Curricular.units.1st.sem..grade.
1                                   0                             0.0
2                                   6                            14.0
3                                   0                             0.0
4                                   6                13.428571428571429
5                                   5                12.333333333333334
  Curricular.units.1st.sem..without.evaluations.
```

```
1                                                0
2                                                0
3                                                0
4                                                0
5                                                0
  Curricular.units.2nd.sem..credited. Curricular.units.2nd.sem..enrolled.
1                                   0                                    0
2                                   0                                    6
3                                   0                                    6
4                                   0                                    6
5                                   0                                    6
  Curricular.units.2nd.sem..evaluations. Curricular.units.2nd.sem..approved.
1                                      0                                   0
2                                      6                                   6
3                                      0                                   0
4                                     10                                   5
5                                      6                                   6
  Curricular.units.2nd.sem..grade.
1                              0.0
2                13.666666666666666
3                              0.0
4                             12.4
5                             13.0
  Curricular.units.2nd.sem..without.evaluations. Unemployment.rate
1                                              0              10.8
2                                              0              13.9
3                                              0              10.8
4                                              0               9.4
5                                              0              13.9
  Inflation.rate   GDP   Target
1            1.4  1.74  Dropout
2           -0.3  0.79 Graduate
3            1.4  1.74  Dropout
4           -0.8 -3.12 Graduate
5           -0.3  0.79 Graduate
```

La tabla se encuentra en formato tabular, esto se puede ver y tambien se comenta en la pagina de descarga

3- Resumen de 5 números de las variables cuantitativas y analizar el mismo.

```r
library(dplyr)


# Se seleciona las variables cuantitaticas
variables_cuantitativas <- select_if(datos, is.numeric)

# Calcular resumen de 5 números para cada variable
#'Vamos a usar sapply para aplicar la funcion fivenum a la base
#'el firenum es una funcion que nos ayuda a calcular el minimo y maximo, los Q1 y Q3, ademas
resumen_5_numeros <- sapply(variables_cuantitativas, fivenum)
resumen_5_numeros <- t(resumen_5_numeros)
#'Para facilitar la lectura vamos a ponerle nosmbres claros a las columnas
colnames(resumen_5_numeros) <- c("Minimo","Q1","Mediana","Q3","Máximo")
print(resumen_5_numeros)
```

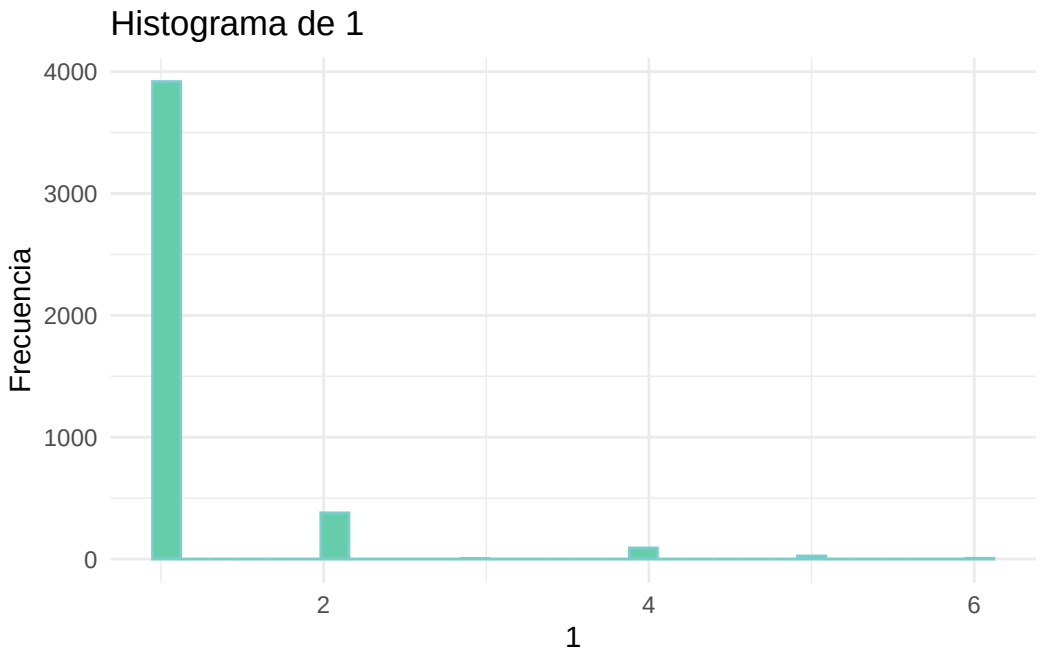
	Minimo	Q1	Mediana	Q3	Máximo
Marital.status	1	1	1	1	6
Application.mode	1	1	17	39	57
Application.order	0	1	1	2	9
Course	33	9085	9238	9556	9991
Daytime.evening.attendance.	0	1	1	1	1
Previous.qualification	1	1	1	1	43
Nacionality	1	1	1	1	109
Mother.s.qualification	1	2	19	37	44
Father.s.qualification	1	3	19	37	44
Mother.s.occupation	0	4	5	9	194
Father.s.occupation	0	4	7	9	195
Displaced	0	0	1	1	1
Educational.special.needs	0	0	0	0	1
Debtor	0	0	0	0	1
Tuition.fees.up.to.date	0	1	1	1	1
Gender	0	0	0	1	1
Scholarship.holder	0	0	0	0	1
Age.at.enrollment	17	19	20	25	70
International	0	0	0	0	1
Curricular.units.1st.sem..credited.	0	0	0	0	20
Curricular.units.1st.sem..enrolled.	0	5	6	7	26
Curricular.units.1st.sem..evaluations.	0	6	8	10	45
Curricular.units.1st.sem..approved.	0	3	5	6	26
Curricular.units.1st.sem..without.evaluations.	0	0	0	0	12
Curricular.units.2nd.sem..credited.	0	0	0	0	19
Curricular.units.2nd.sem..enrolled.	0	5	6	7	23

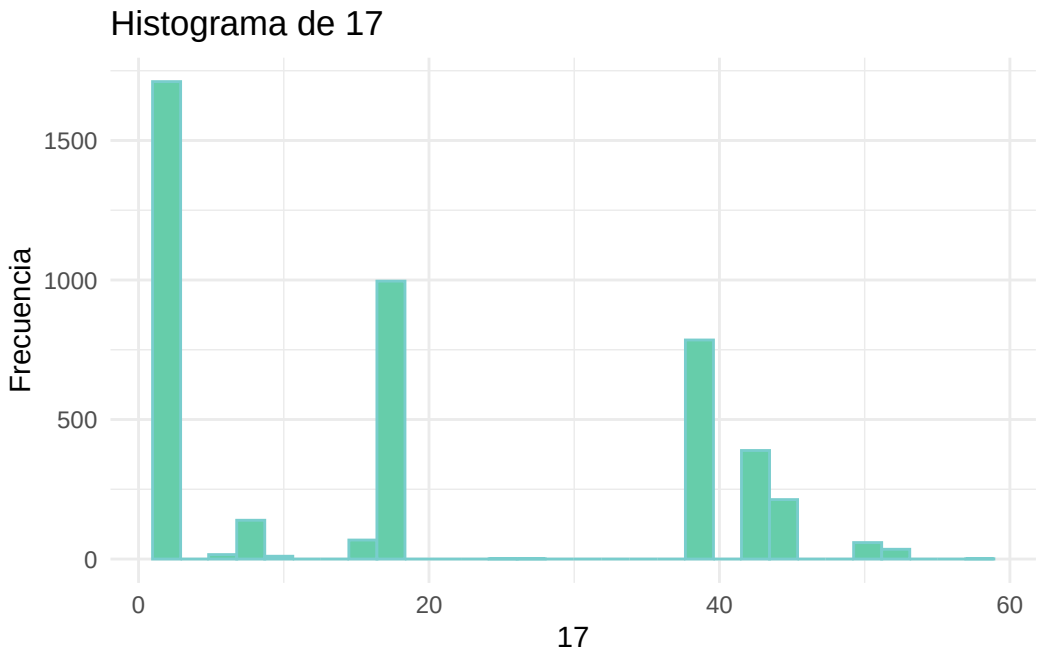
```
Curricular.units.2nd.sem..evaluations.              0    6       8   10      33
Curricular.units.2nd.sem..approved.                 0    2       5    6      20
Curricular.units.2nd.sem..without.evaluations.      0    0       0    0      12
```

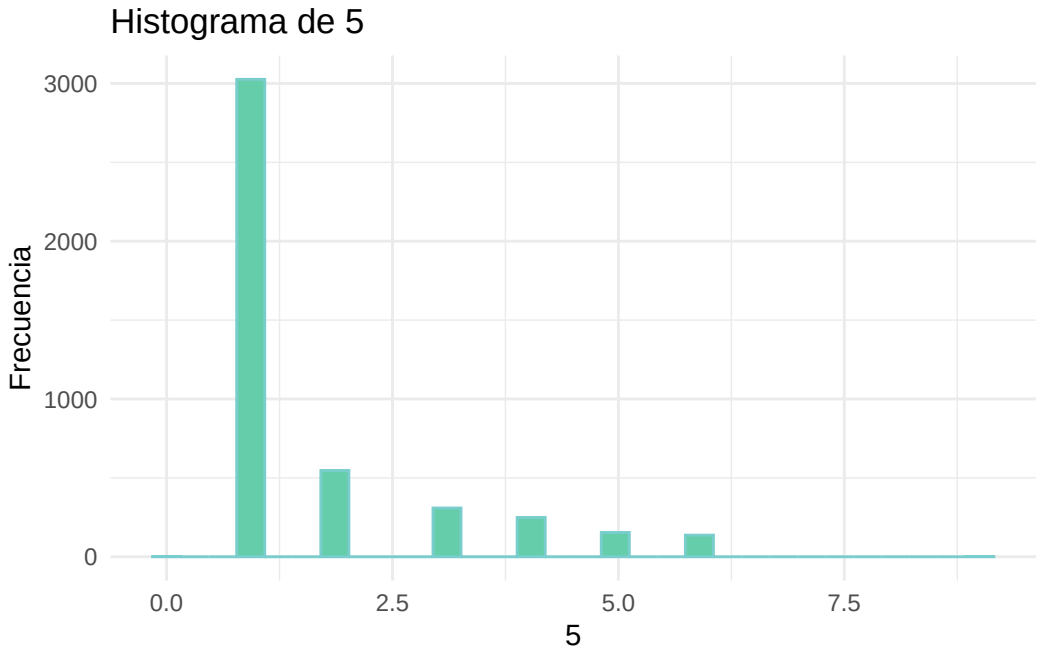
4- Hacer al menos un gráfico que describa la distribución para cada una de las variables cuantitativas.

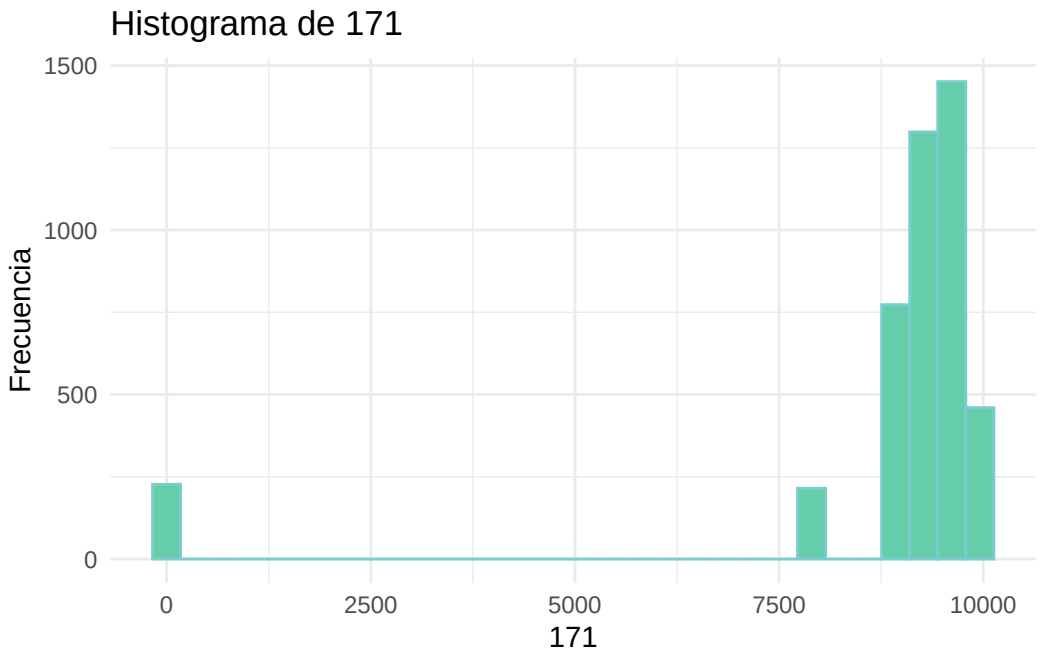
```
#str(datos)

for (var in variables_cuantitativas){
  h <- ggplot(datos, aes_string(x=var))+
    geom_histogram(fill = "#66CDAA", color = "#79CDCD", bins = 30) +
     labs(title = paste("Histograma de", var), x = var, y = "Frecuencia") +
    theme_minimal()
  print(h)
}
```

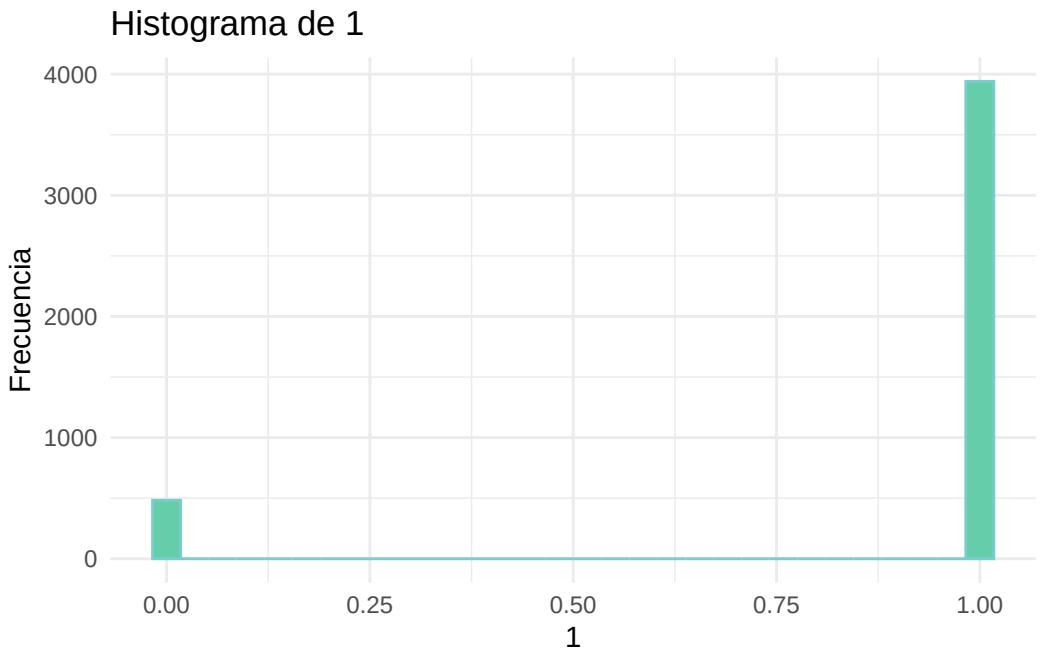
```
Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
i Please use tidy evaluation idioms with `aes()`.
i See also `vignette("ggplot2-in-packages")` for more information.
```

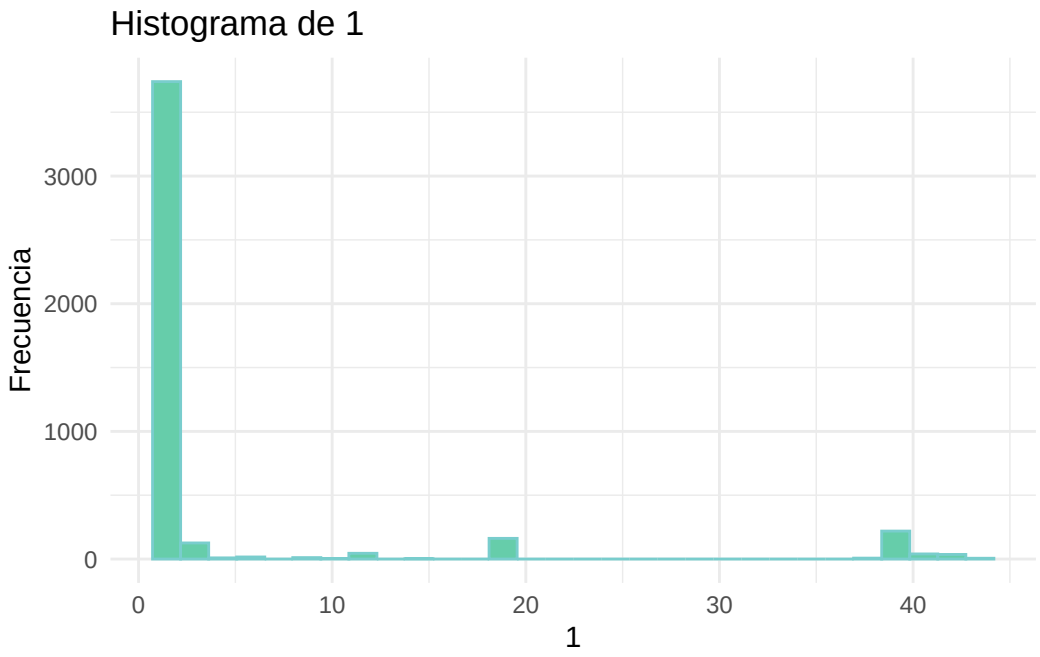


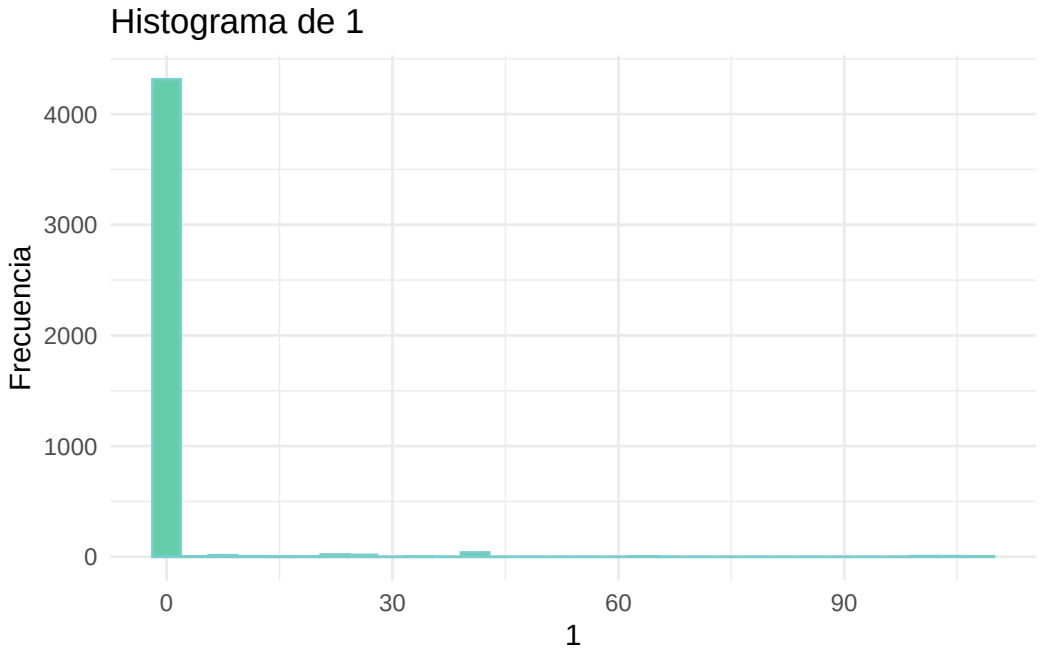

Histograma de 17
Frecuencia
1500
1000
500
0
0
20
40
60
17

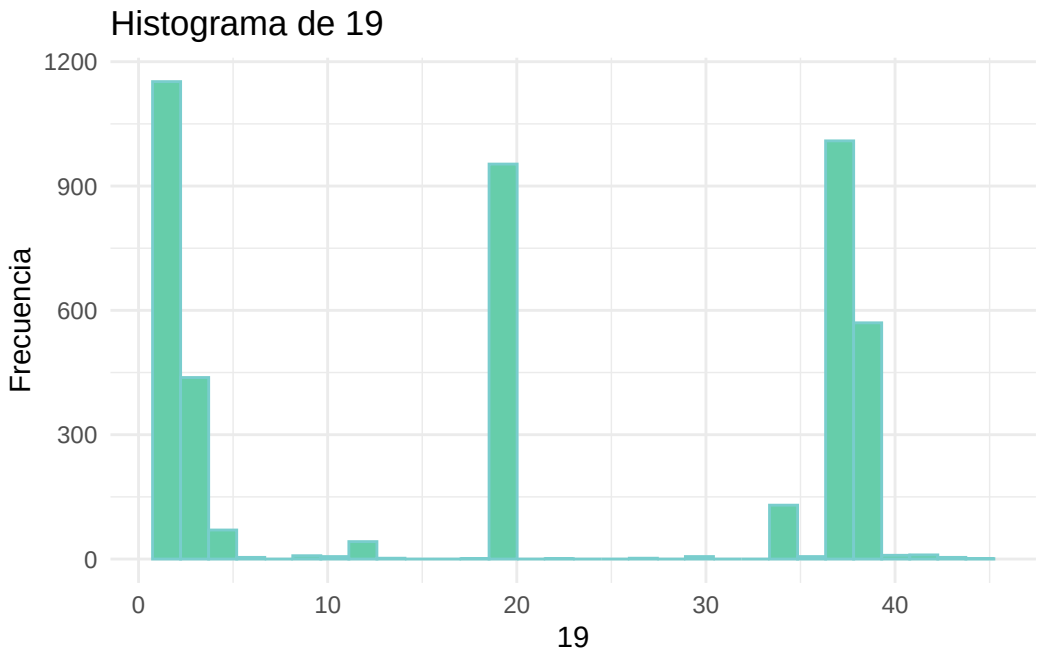

Histograma de 5
Frecuencia
3000
2000
1000
0
0.0
2.5
5.0
7.5
5

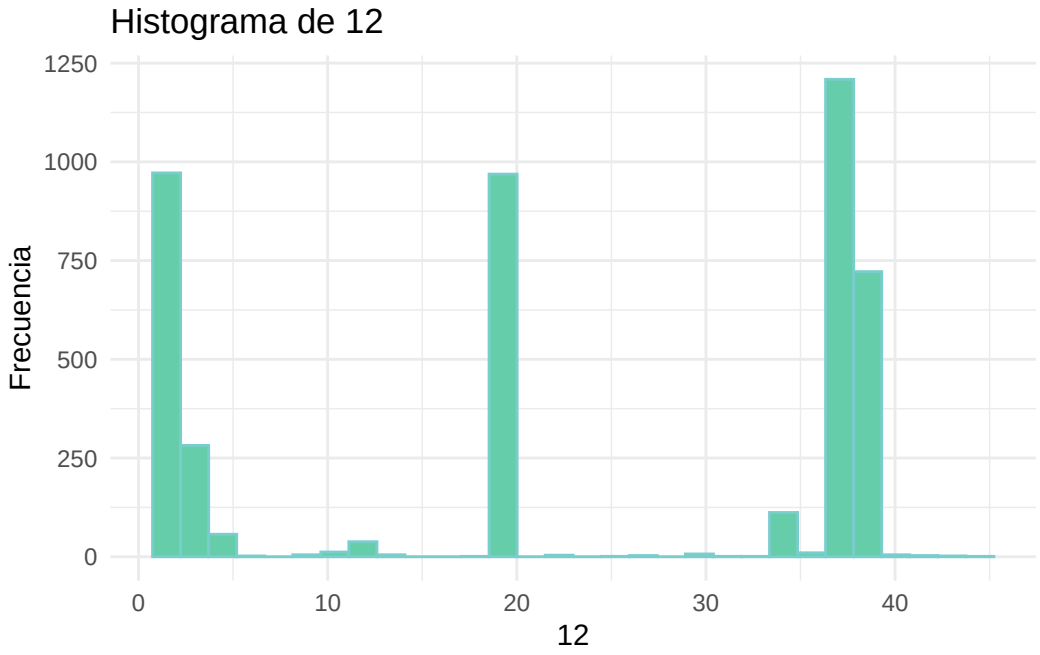

Histograma de 171
Frecuencia
1500
1000
500
0
0
2500
5000
7500
10000
171

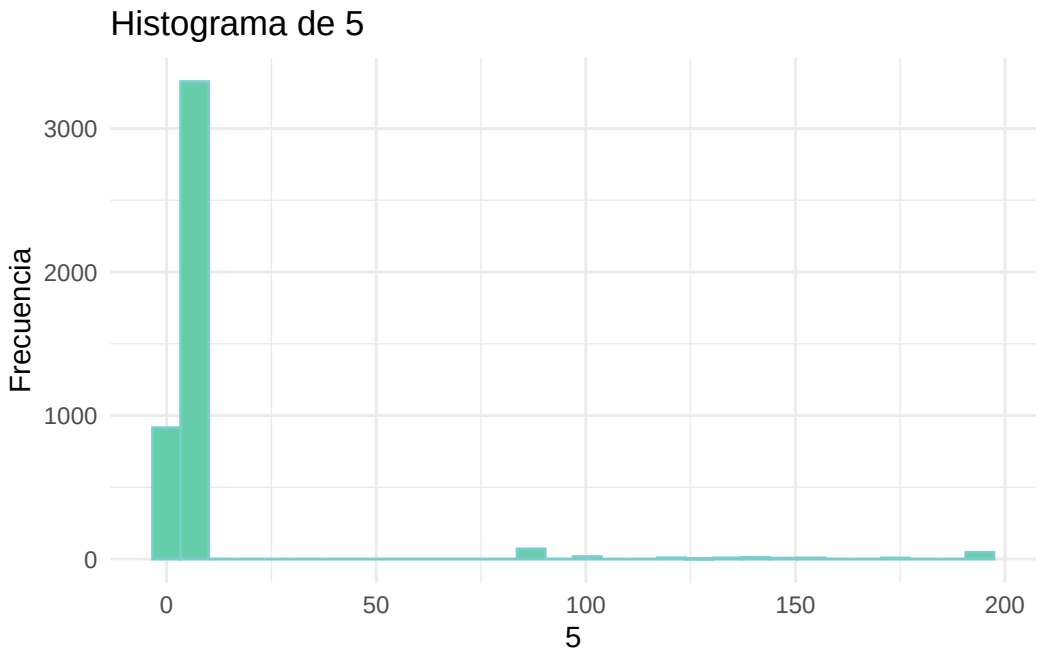

Histograma de 1
Frecuencia
4000
3000
2000
1000
0
0.00
0.25
0.50
0.75
1.00
1

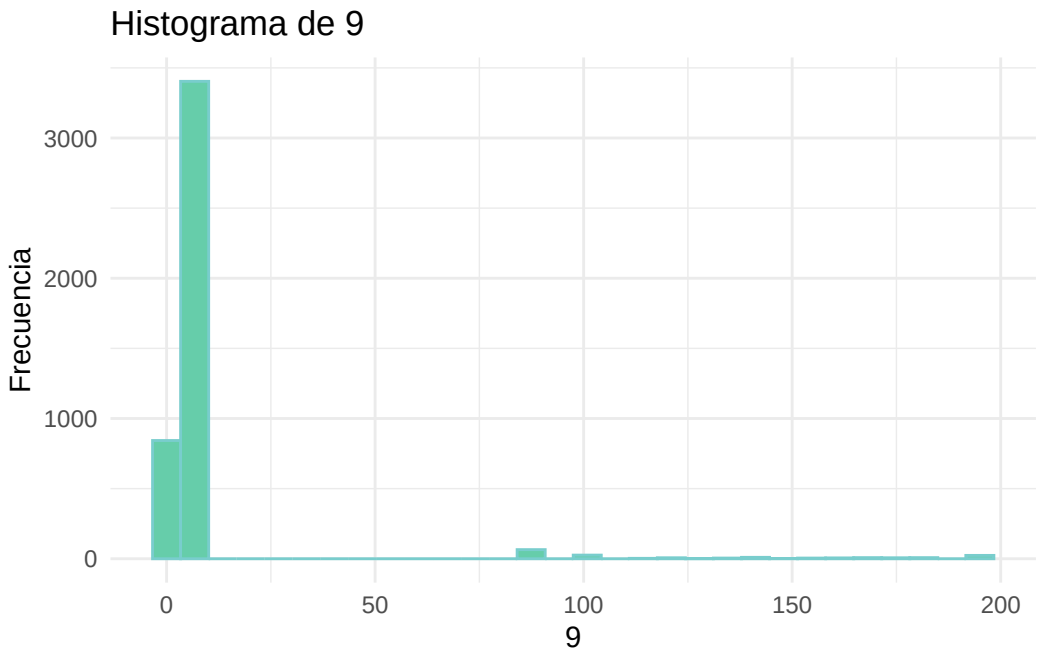

Histograma de 1
Frecuencia
3000
2000
1000
0
0
10
20
30
40
1

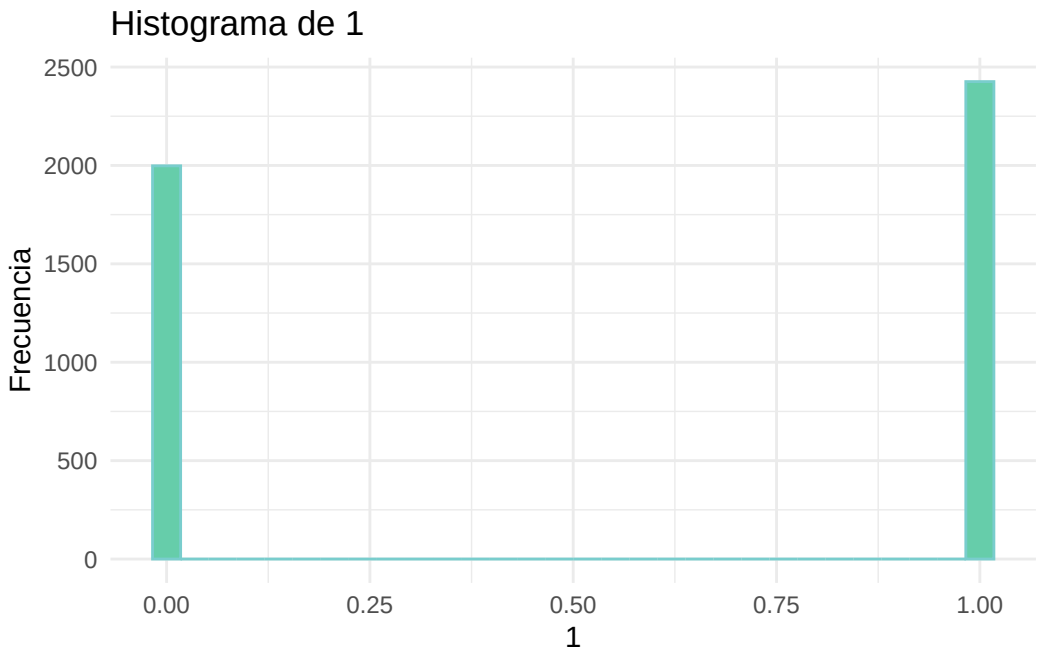

Histograma de 1
Frecuencia
4000
3000
2000
1000
0
0
30
60
90
1

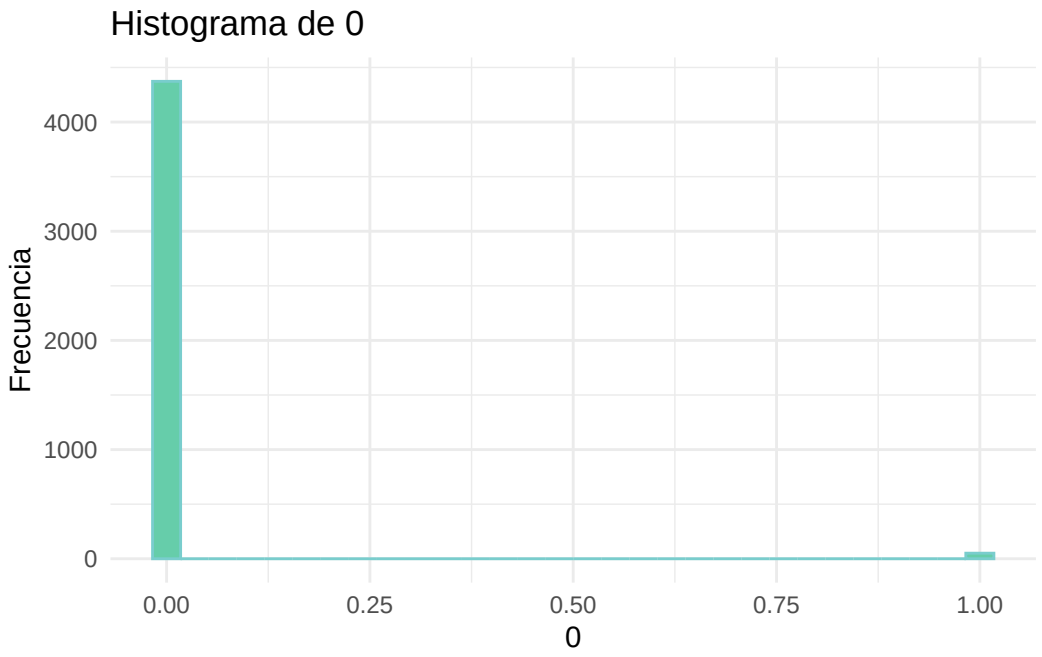

Histograma de 19
Frecuencia
1200
900
600
300
0
0
10
20
30
40
19

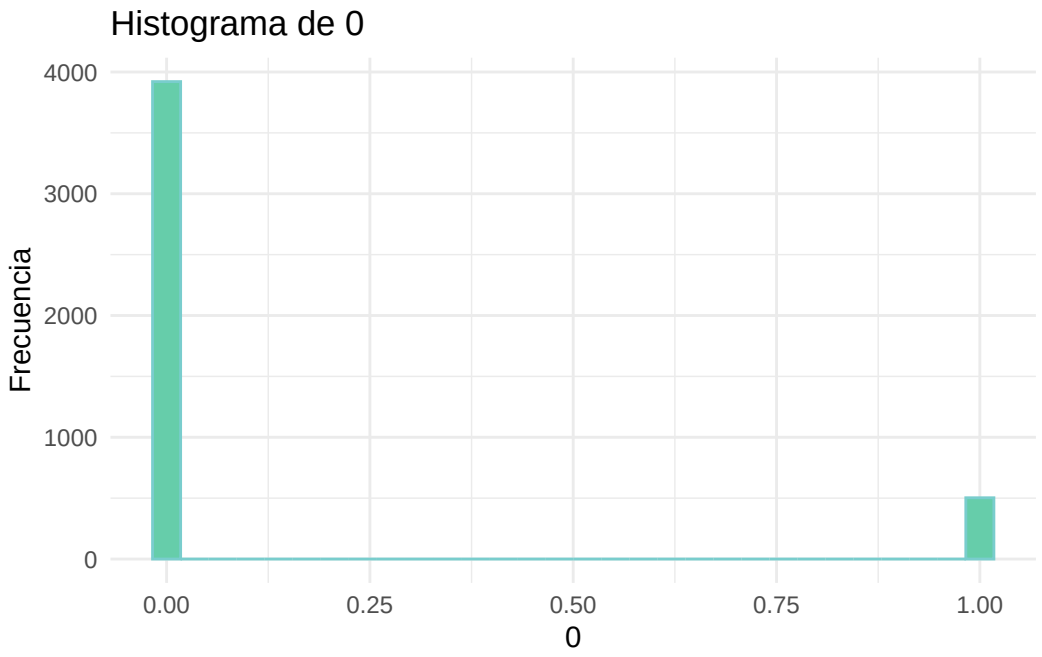

Histograma de 12
Frecuencia
1250
1000
750
500
250
0
0
10
20
30
40
12

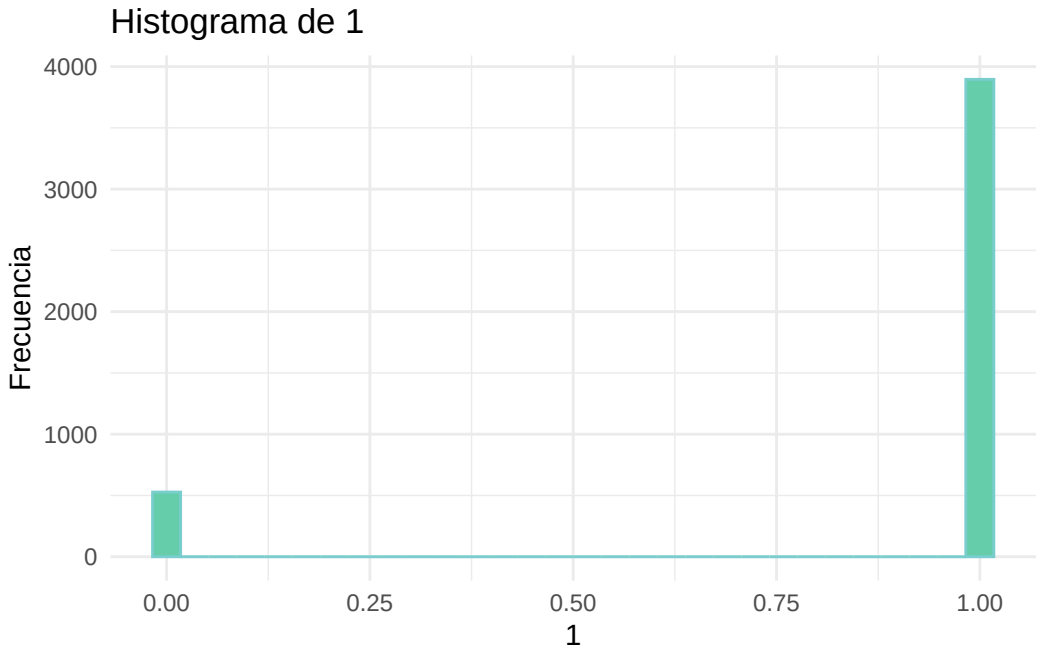

Histograma de 5
Frecuencia
3000
2000
1000
0
0
50
100
150
200
5

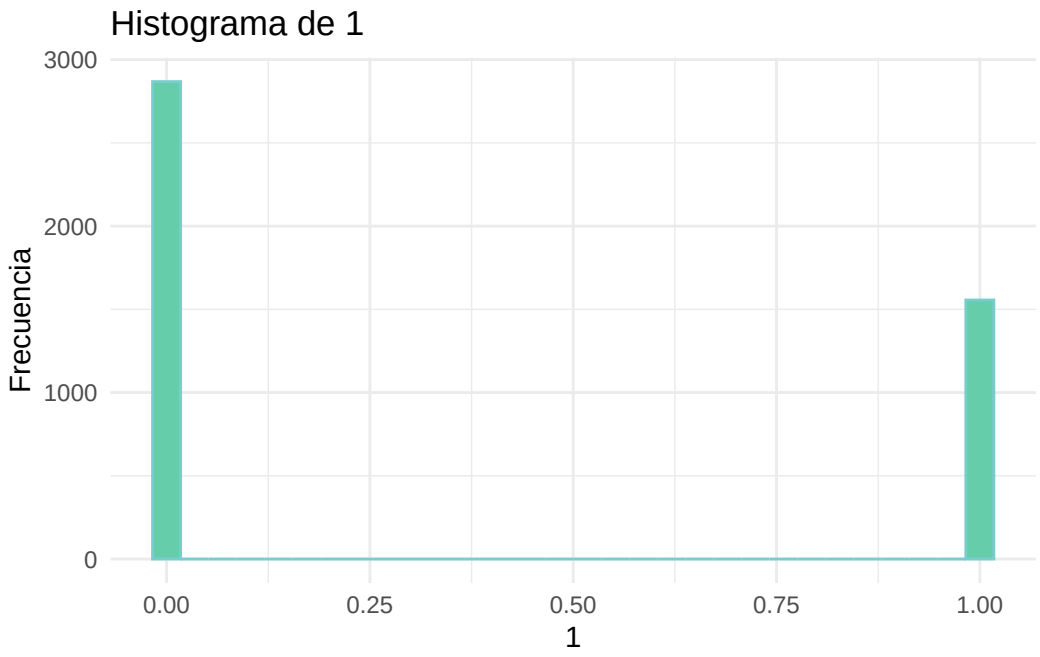

Histograma de 9
Frecuencia
3000
2000
1000
0
0
50
100
150
200
9

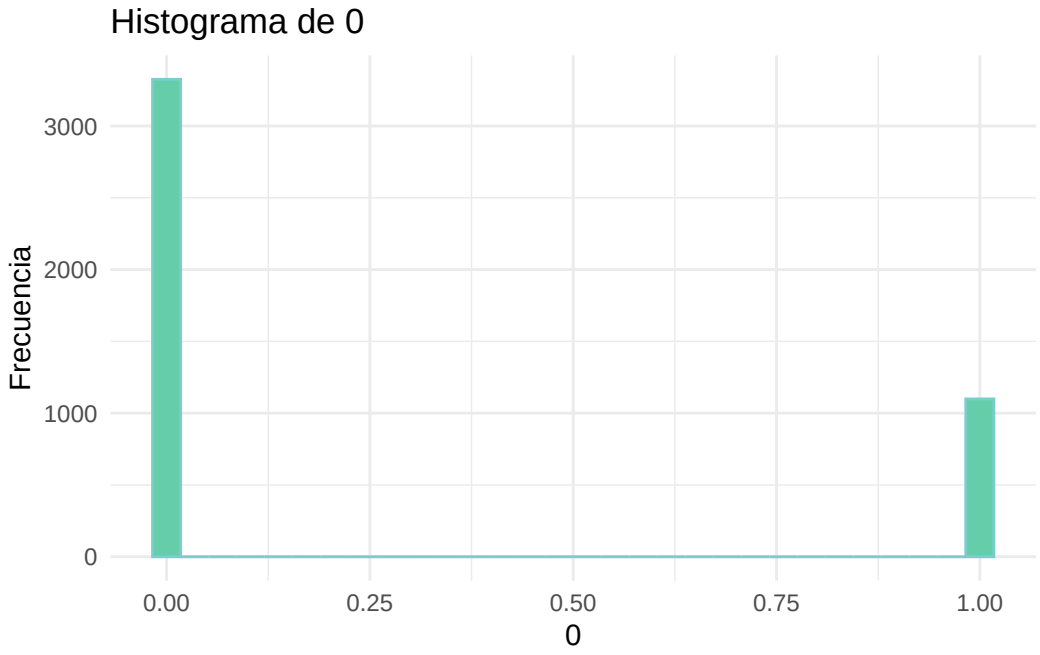

Histograma de 1
Frecuencia
2500
2000
1500
1000
500
0
0.00
0.25
0.50
0.75
1.00
1

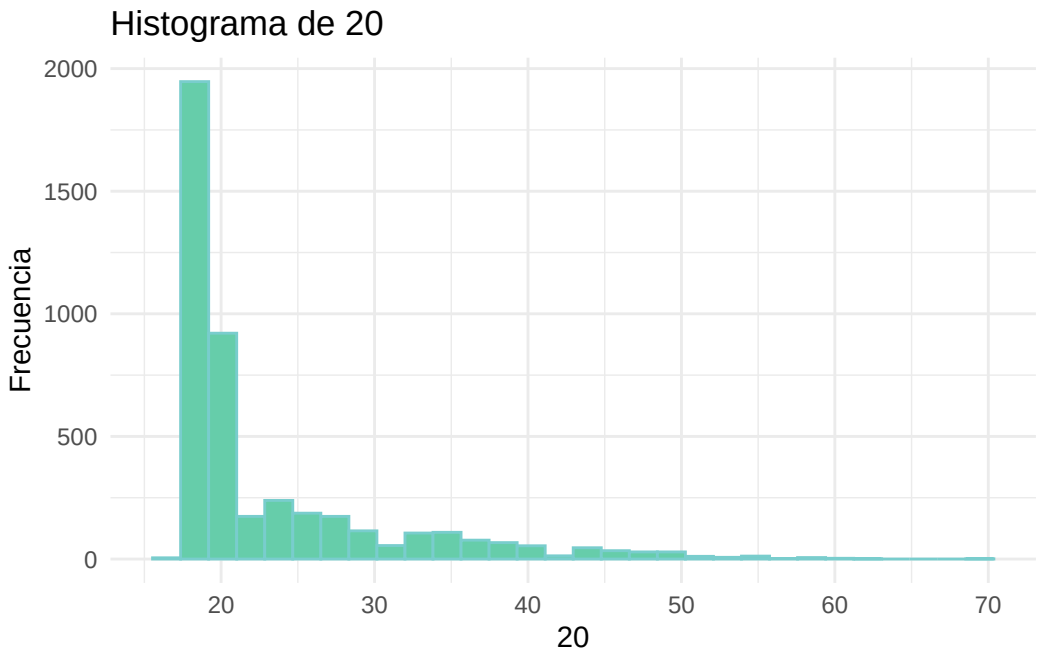

Histograma de 0
Frecuencia
4000
3000
2000
1000
0
0.00
0.25
0.50
0.75
1.00
0

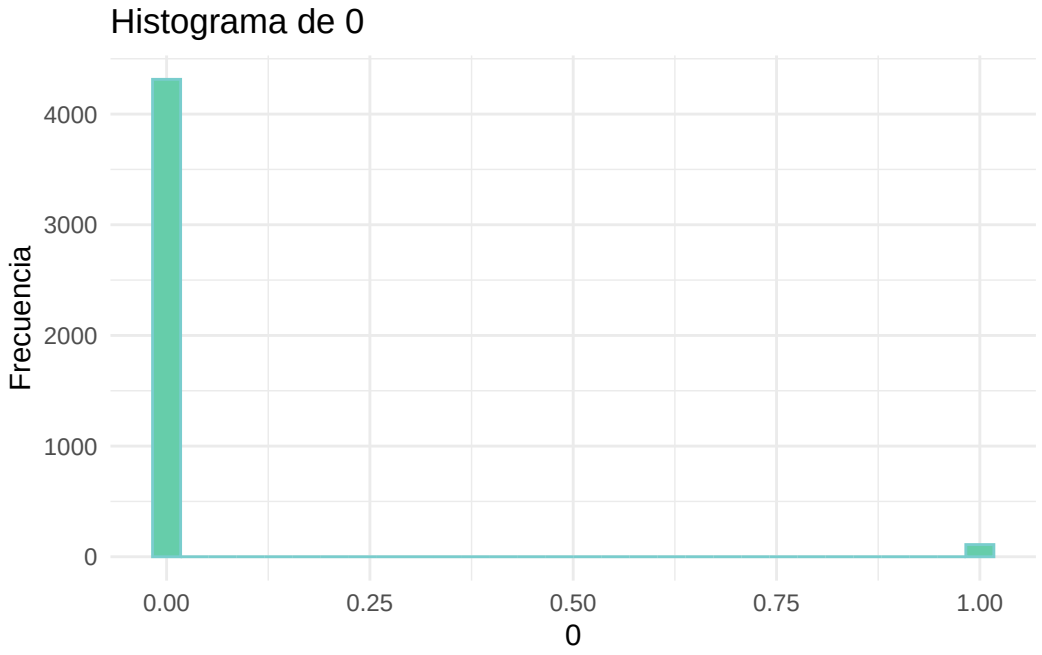

Histograma de 0
Frecuencia
4000
3000
2000
1000
0
0.00
0.25
0.50
0.75
1.00
0

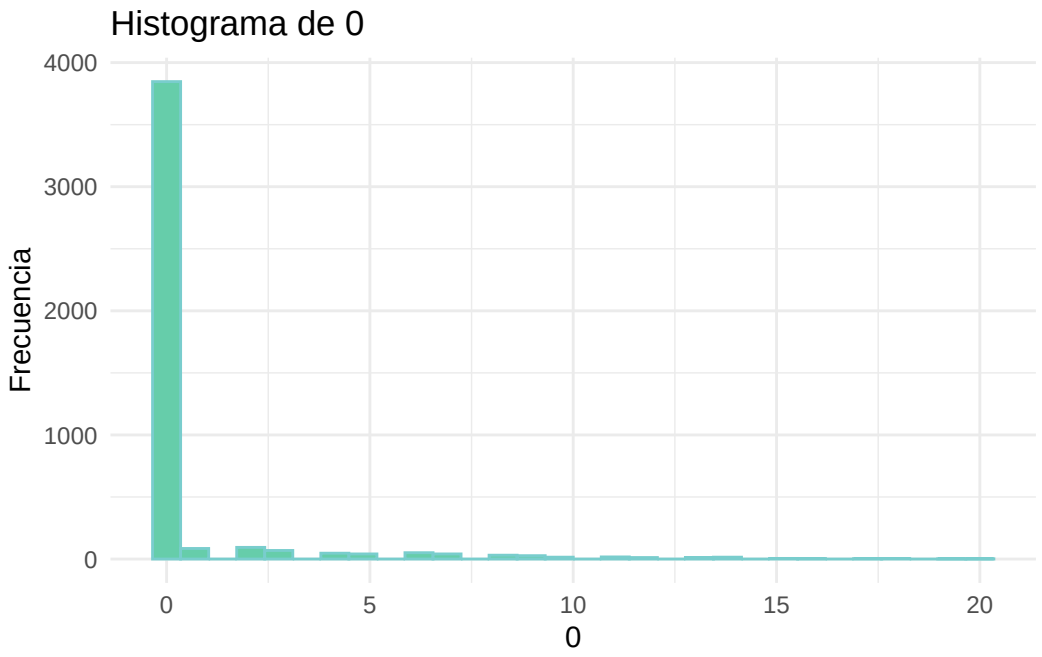

Histograma de 1
Frecuencia
4000
3000
2000
1000
0
0.00
0.25
0.50
0.75
1.00
1

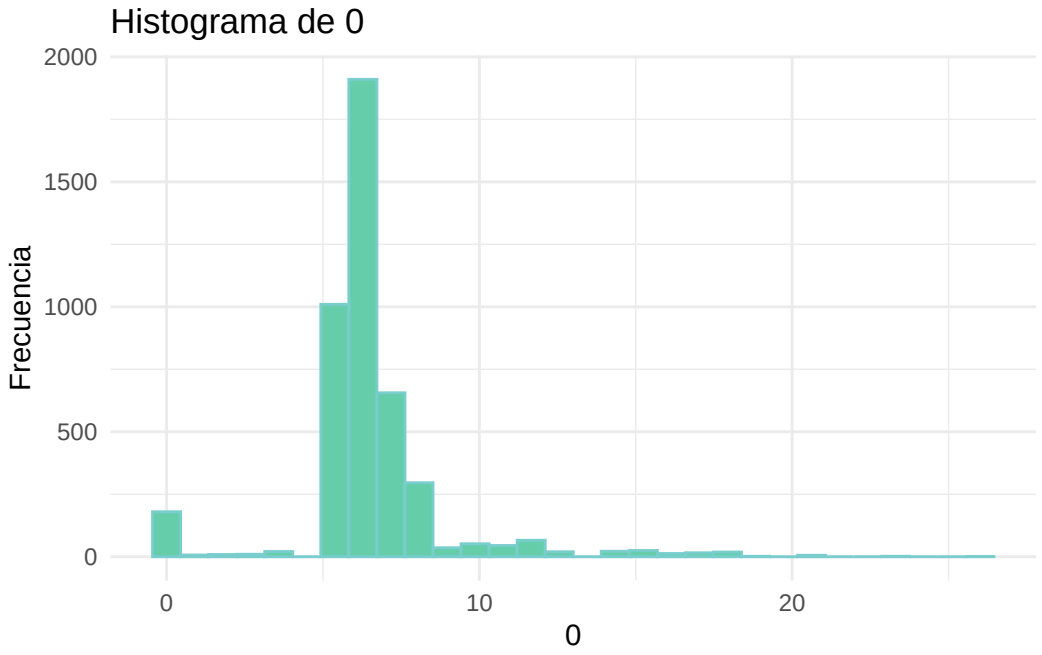

Histograma de 1
Frecuencia
3000
2000
1000
0
0.00
0.25
0.50
0.75
1.00
1

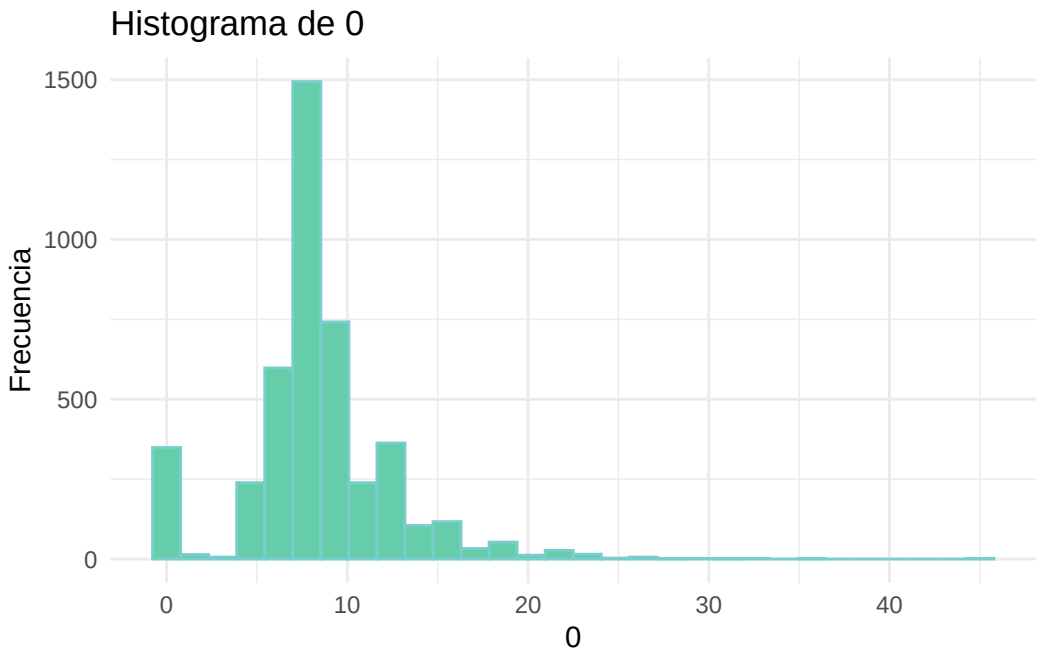

Histograma de 0
Frecuencia
3000
2000
1000
0
0.00
0.25
0.50
0.75
1.00
0

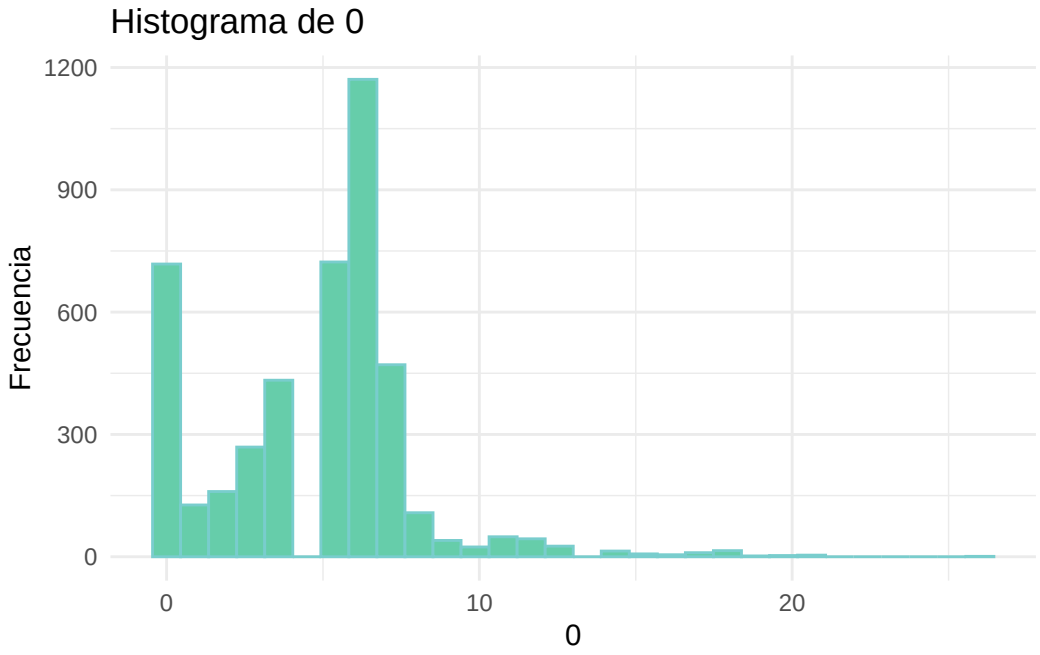

Histograma de 20
Frecuencia
2000
1500
1000
500
0
20
30
40
50
60
70
20

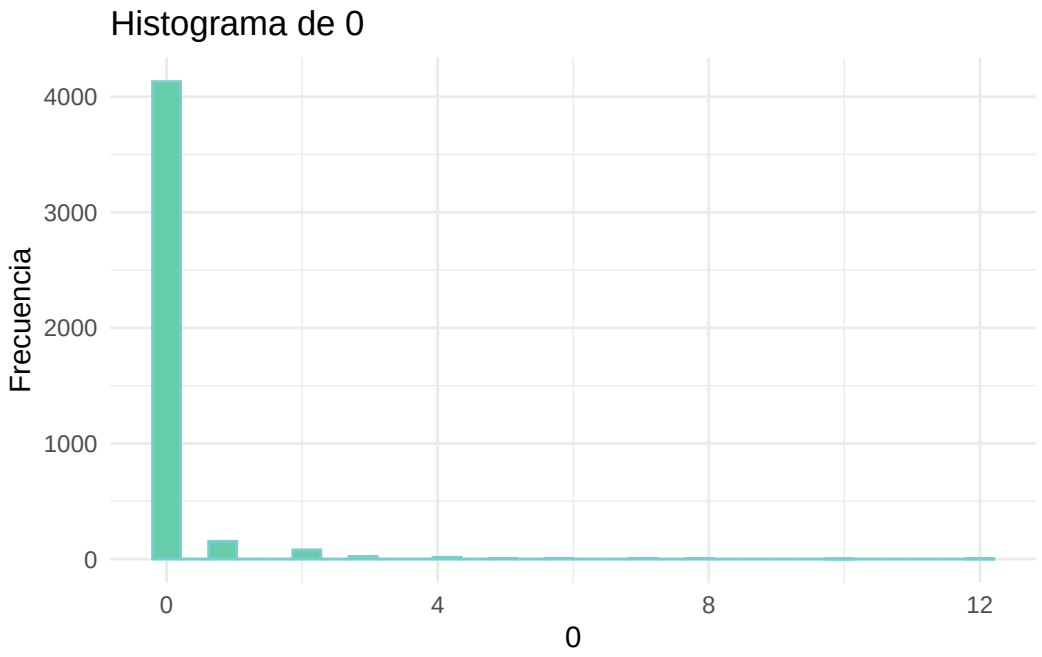

Histograma de 0
Frecuencia
4000
3000
2000
1000
0
0.00
0.25
0.50
0.75
1.00
0

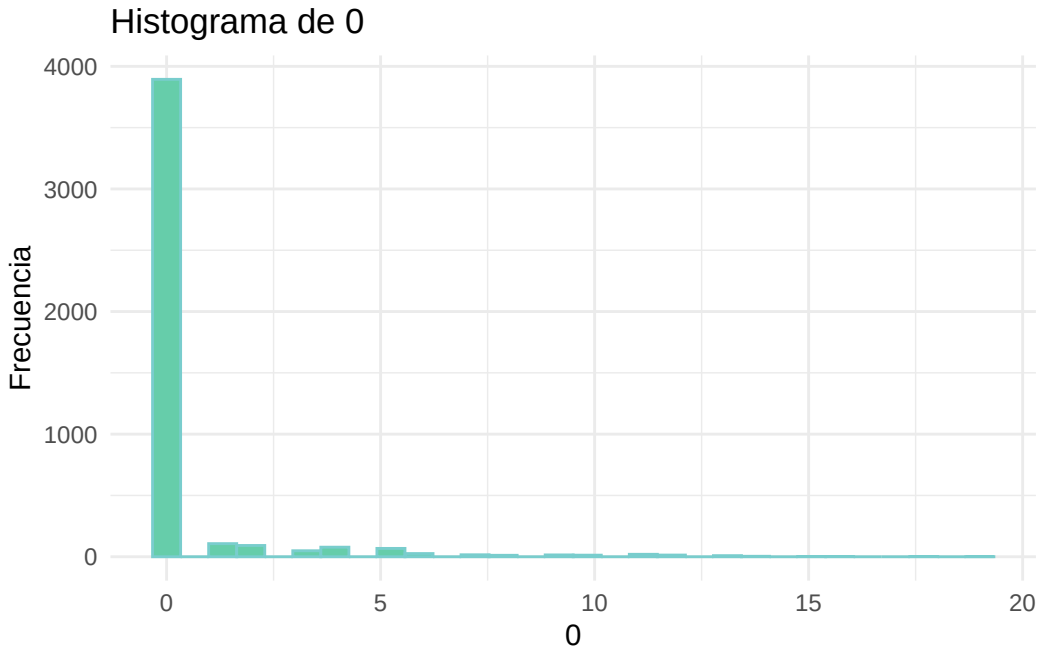

Histograma de 0
Frecuencia
4000
3000
2000
1000
0
0
5
10
15
20
0

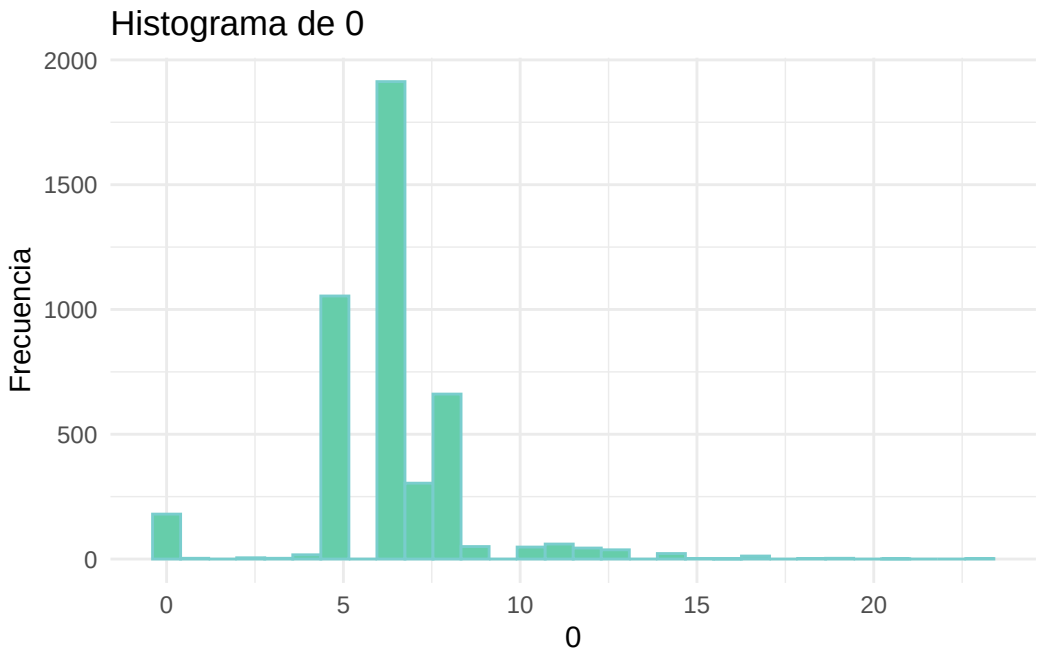

Histograma de 0
Frecuencia
2000
1500
1000
500
0
0
10
20
0

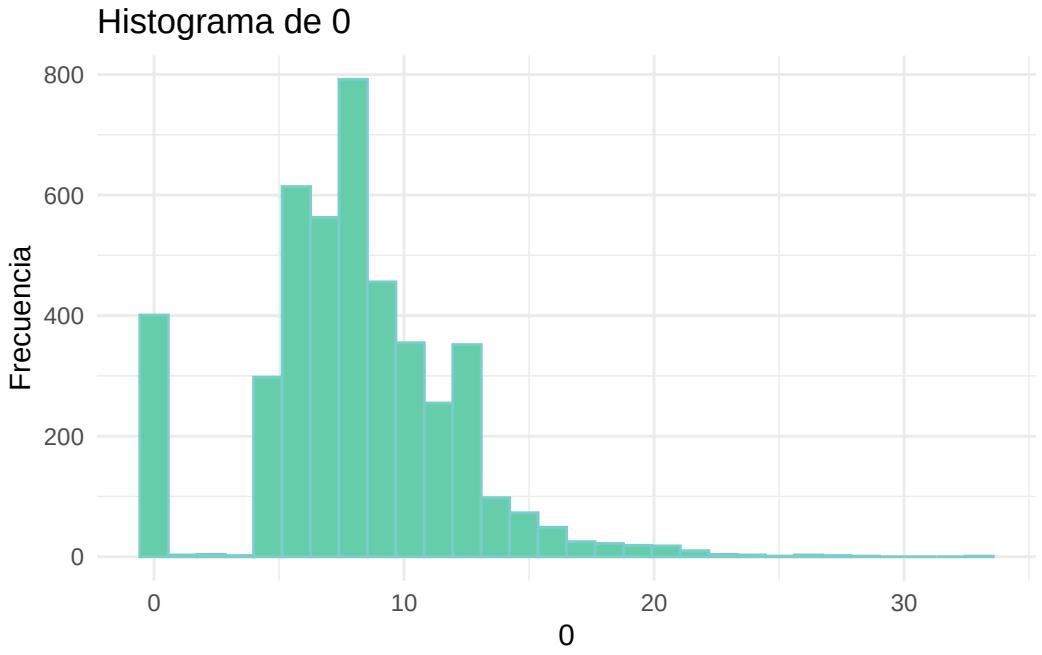

Histograma de 0
Frecuencia
1500
1000
500
0
0
10
20
30
40
0

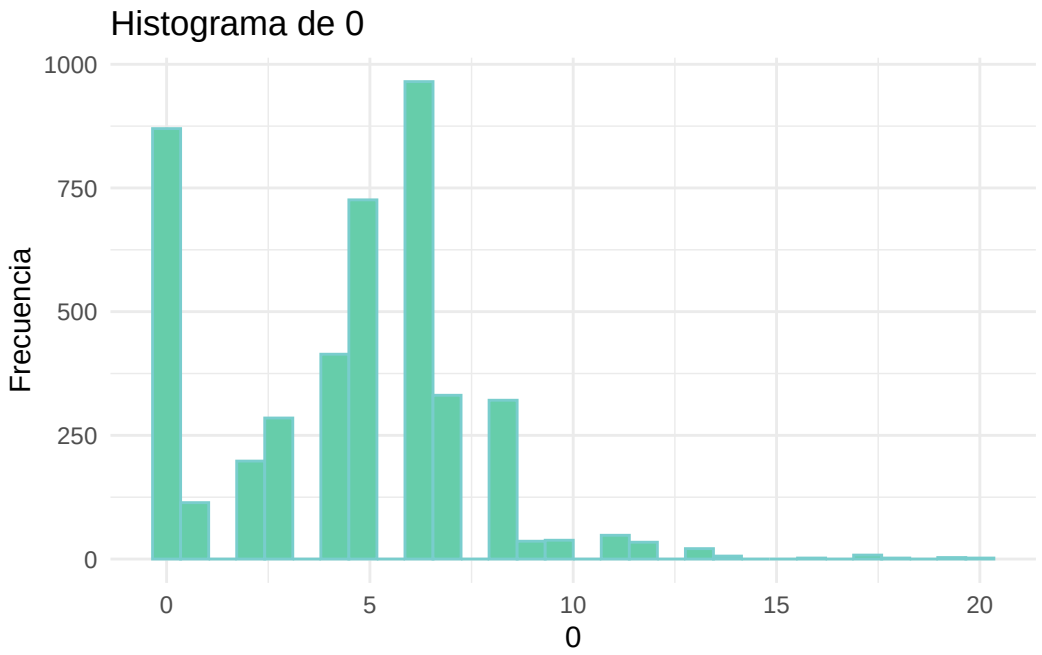

Histograma de 0
Frecuencia
1200
900
600
300
0
0
10
20
0

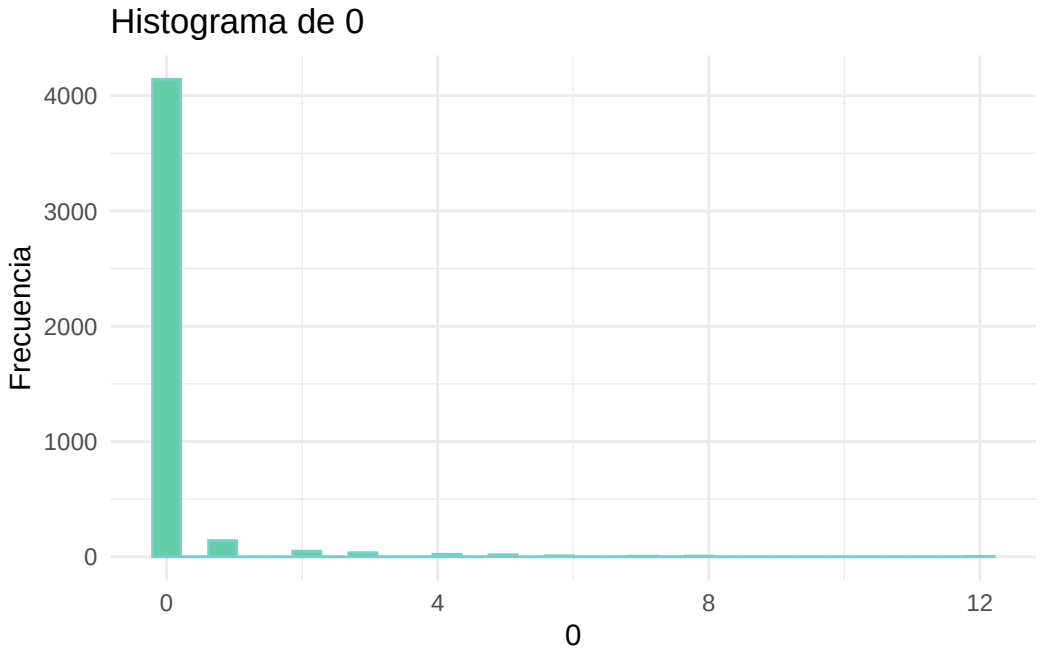

Histograma de 0
Frecuencia
4000
3000
2000
1000
0
0
4
8
12
0


Histograma de 0
Frecuencia
4000
3000
2000
1000
0
0
5
10
15
20
0


Histograma de 0
2000
1500
1000
500
0
Frecuencia
0
5
10
15
20
0


Histograma de 0
800
600
400
200
0
Frecuencia
0
10
20
30
0





5- Hacer al menos dos gráficos que describan la relación entre las variables.

6- Hacer al menos un gráfico que muestre la distribución de las variables categóricas.

7- Identificar valores faltantes y posibles outliers.

8- Investigar técnicas que permitan subsanar los valores perdidos y outliers.

Bibliografia

https://www.maximaformacion.es/blog-dat/como-describir-tus-datos-en-r-paso-1/

https://rpubs.com/Elyn1017/Aunivariado_Vcuantitativas_CasoMedicos