

BITÁCORA 2 - GRUPO 2

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.

Características 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)
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.

```

library(dplyr)
# Se selecciona las variables cuantitativas
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, ad
resumen_5_numeros <- sapply(variables_cuantitativas, fivenum)
#para no tener problema trasponemos a resumen_5_numeros
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

La tabla muestra el resumen de 5 numeros de las variables cuantitativas de nuestra base de datos. en la tabla se puede ver que la mayoría de las variables tienen un minimo de en 0, la forma en la que se evalua nuestra variables es de una manera difereente ya que estas van con diferentes rangos par asignarles puede que sea continio o que se llegue a saltar numeros, pasa de 33 o 53 y cosas así.

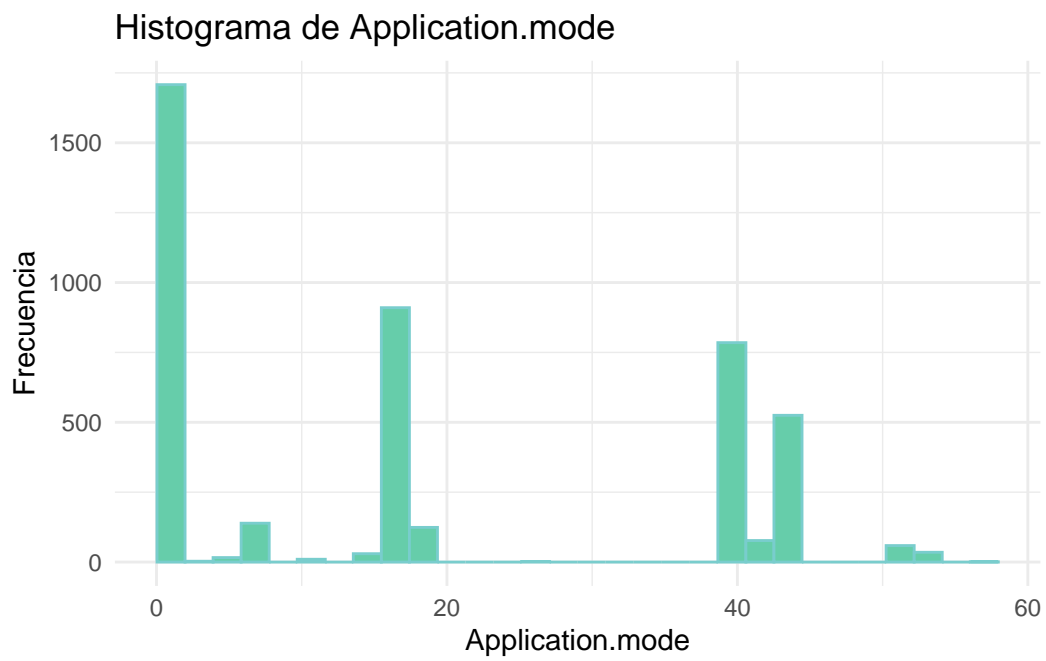
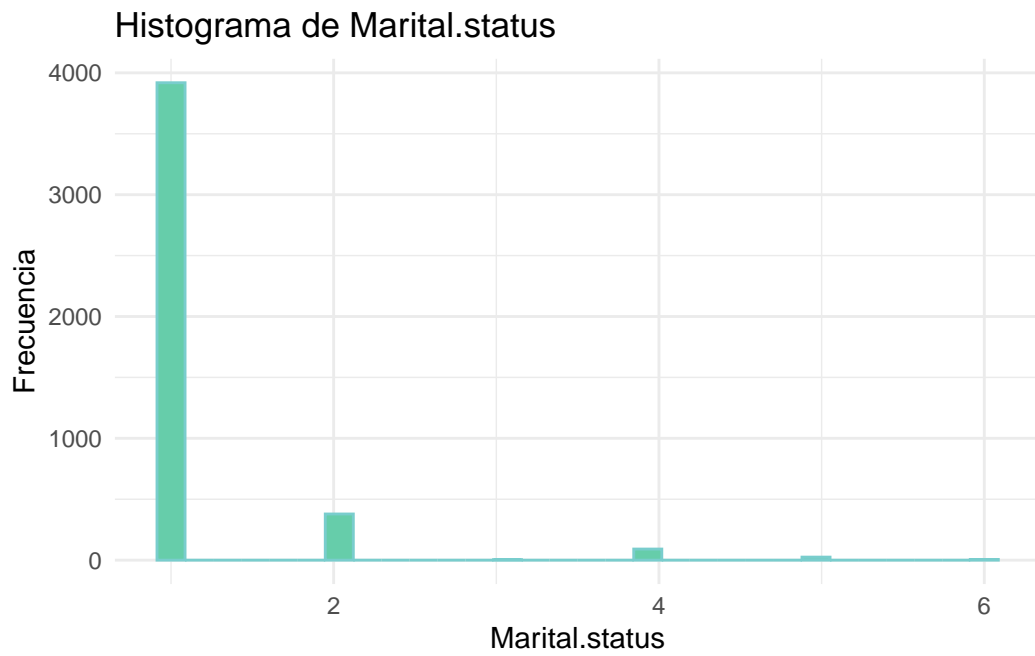
4- Hacer al menos un gráfico que describa la distribución para cada una de las variables cuantitativas.

```
#str(datos)

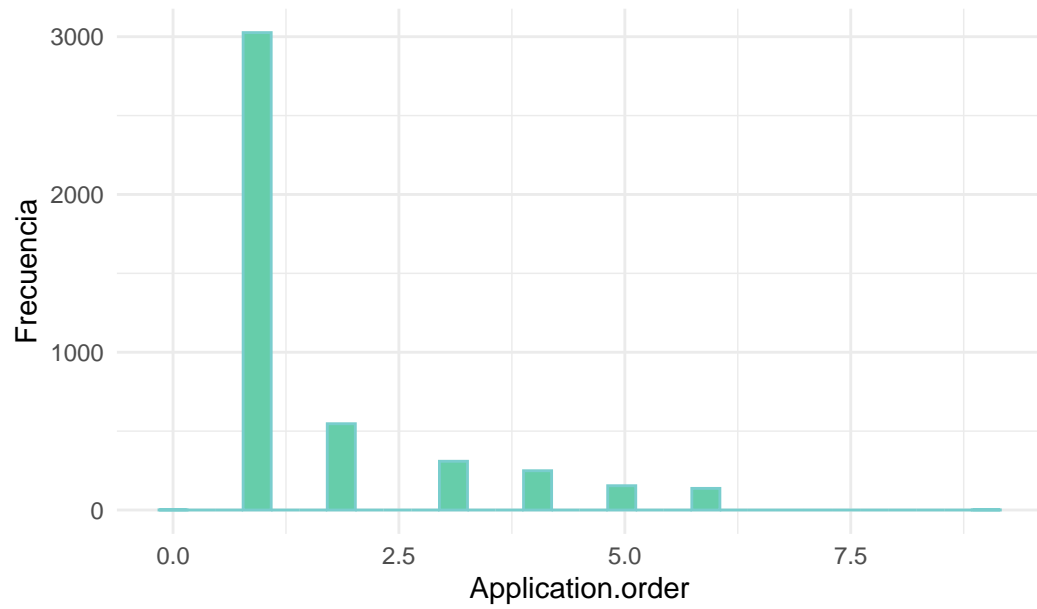
variables_cuantitativas <- names(select(datos, where(is.numeric)))

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

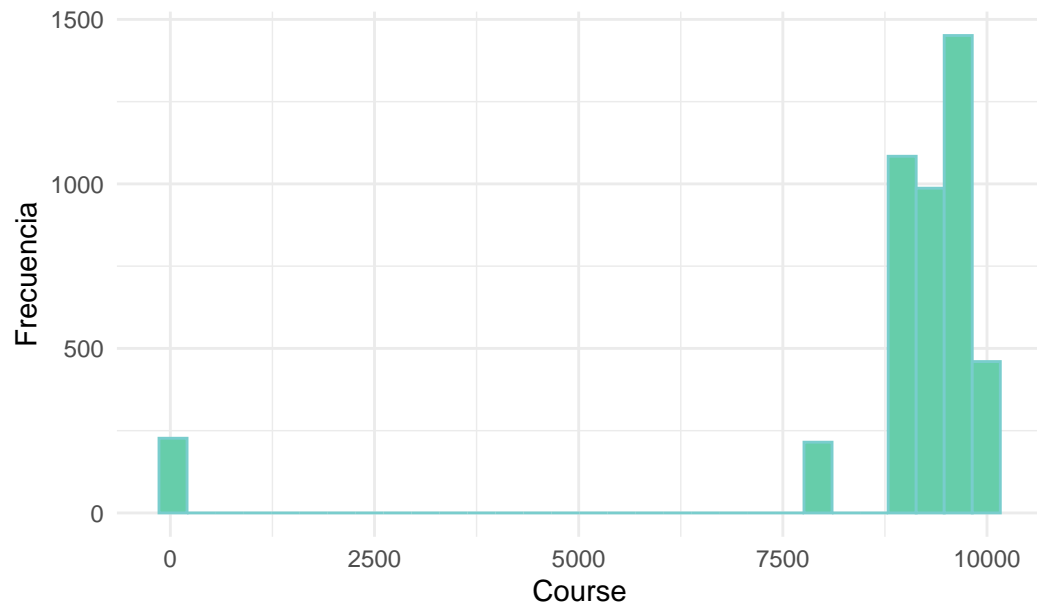
}



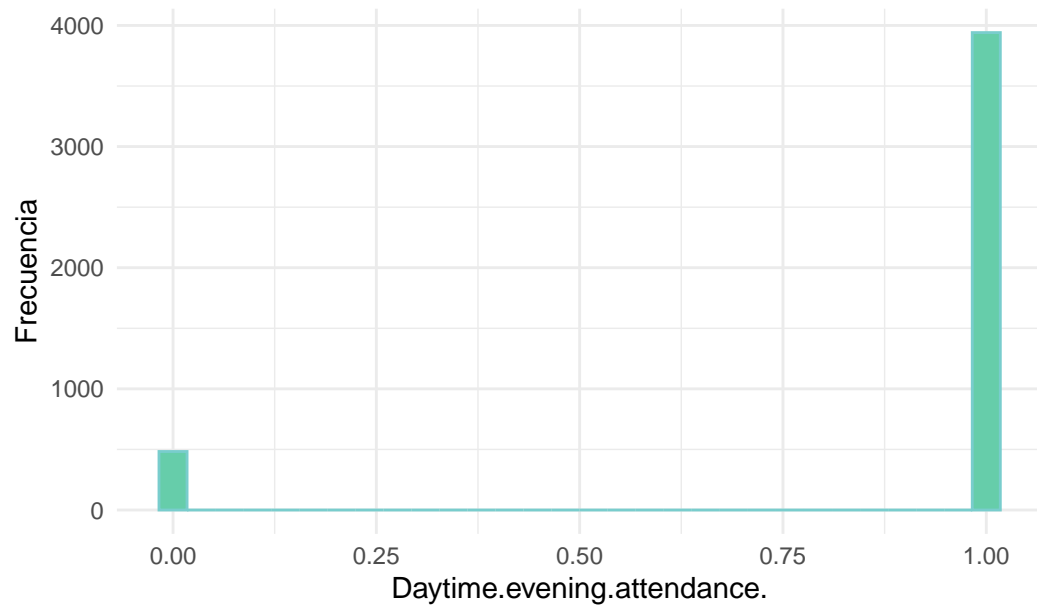
Histograma de Application.order



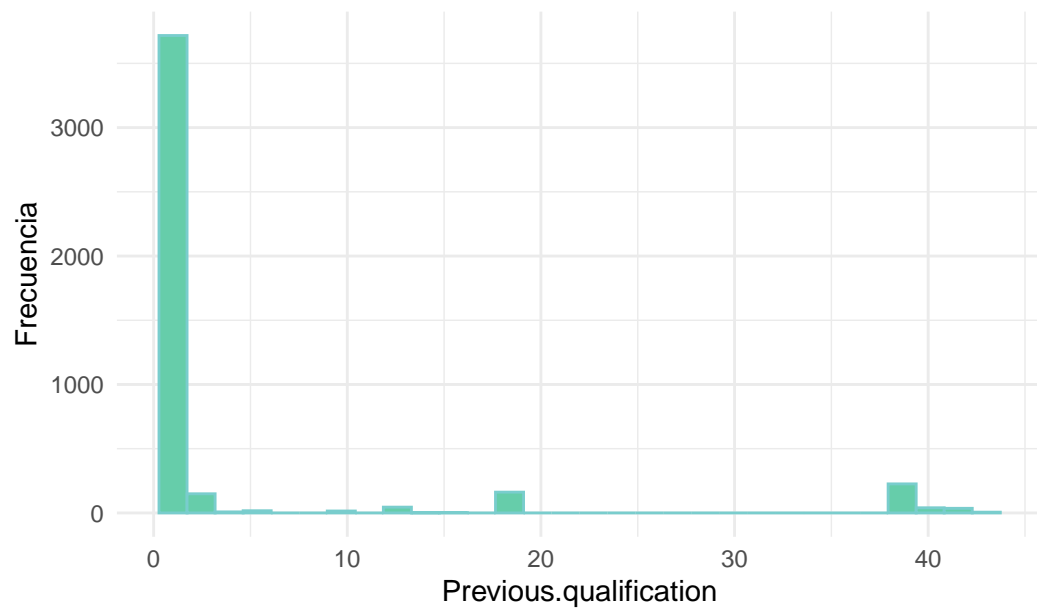
Histograma de Course

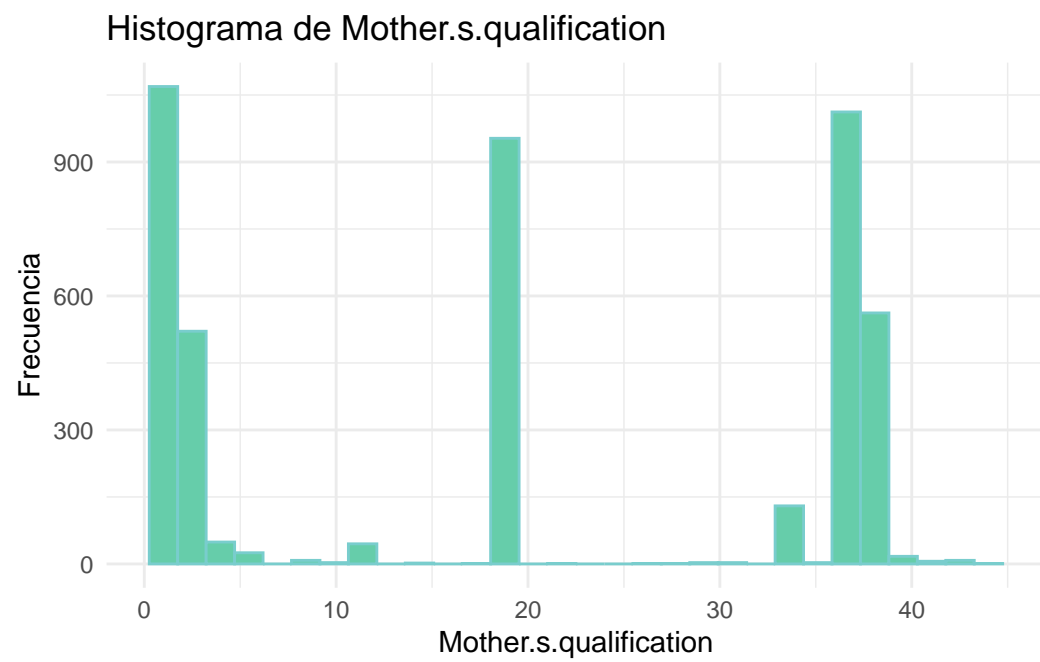
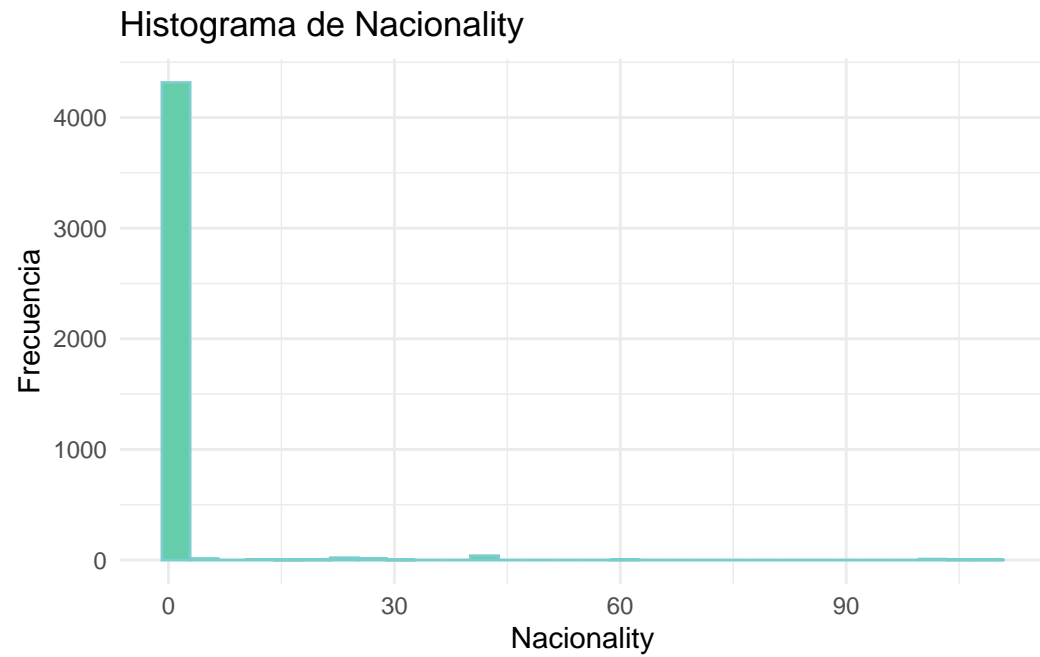


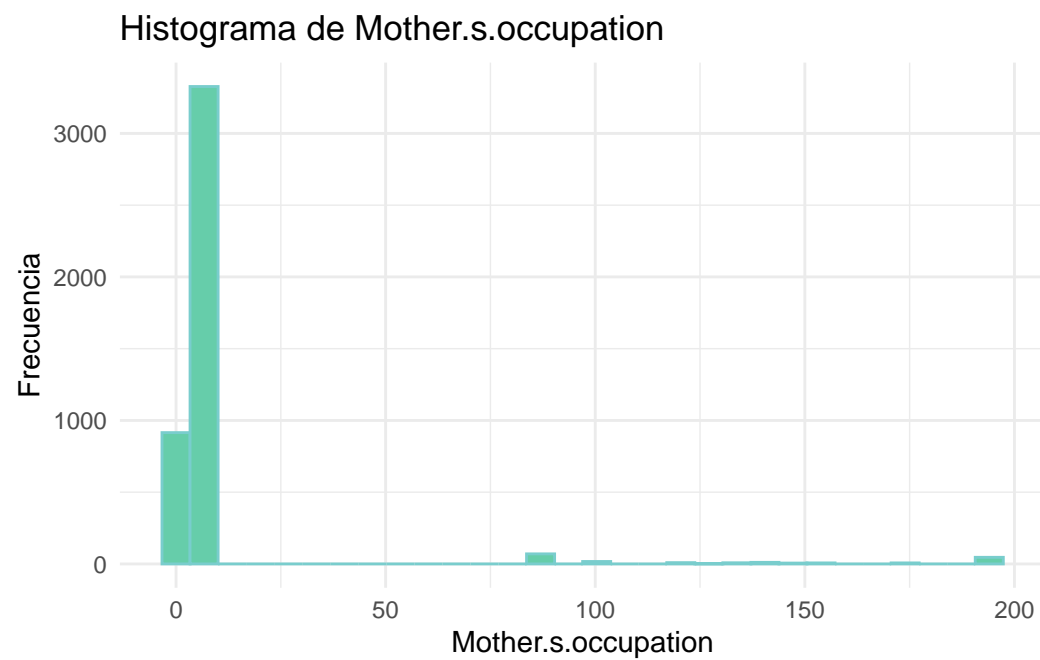
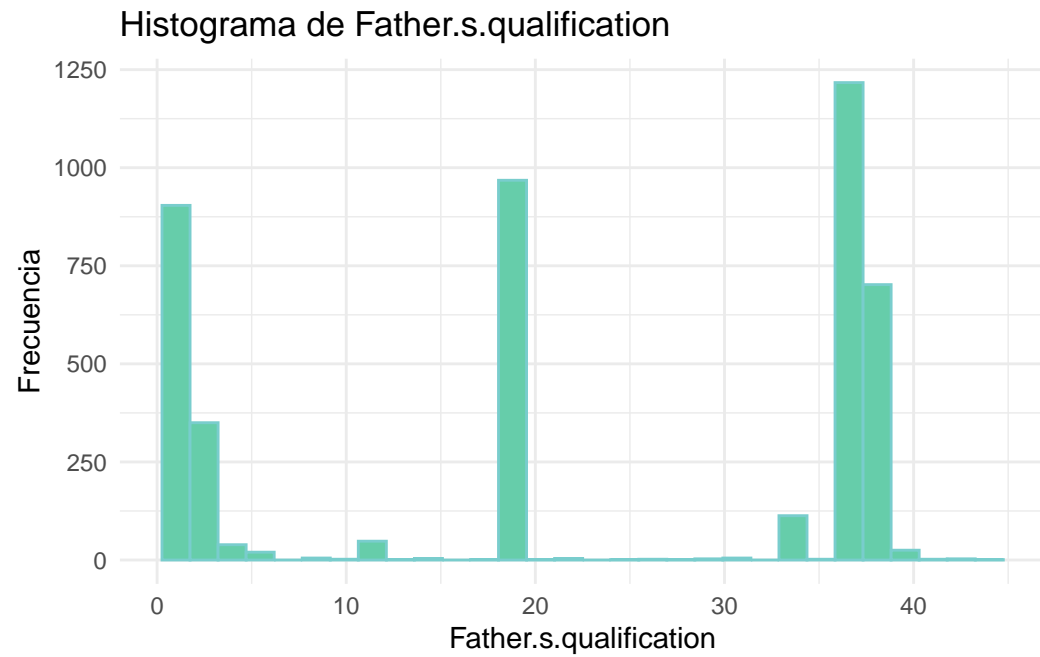
Histograma de Daytime.evening.attendance.

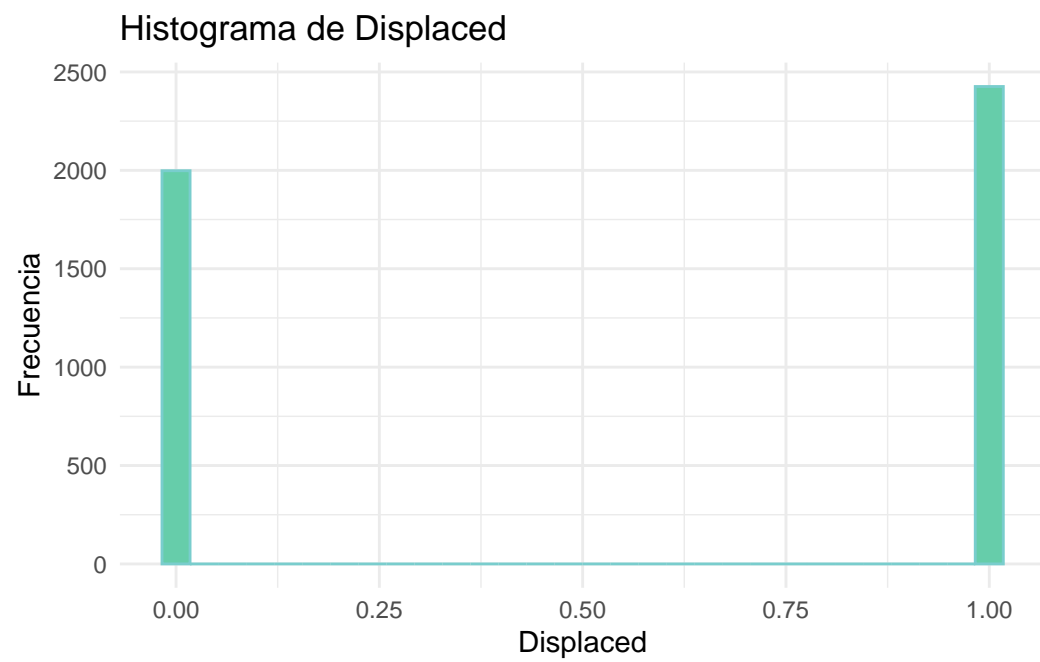
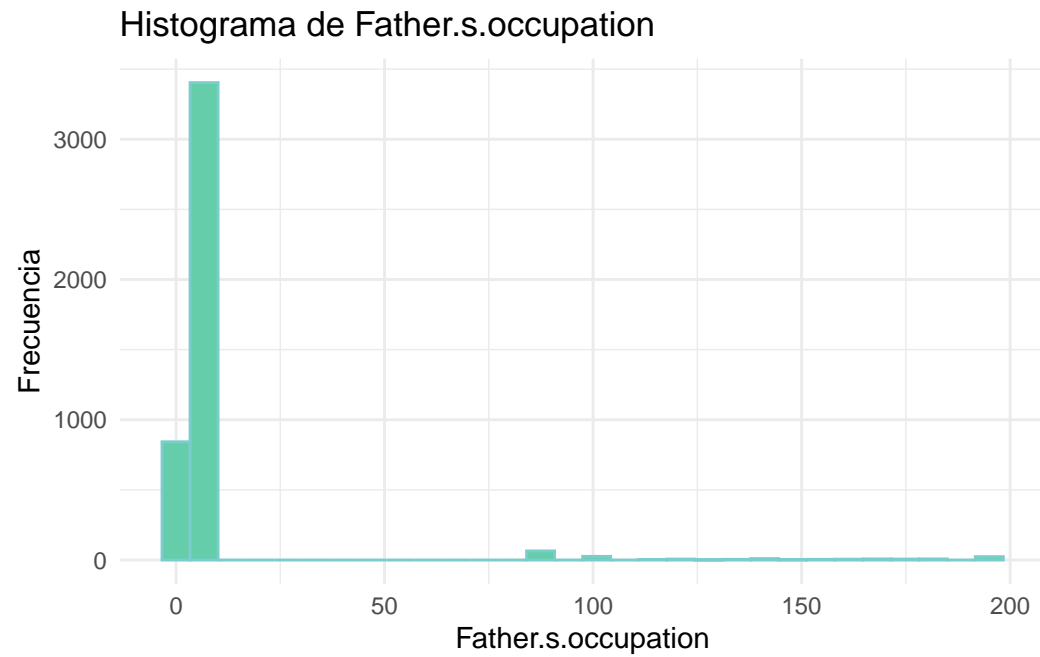


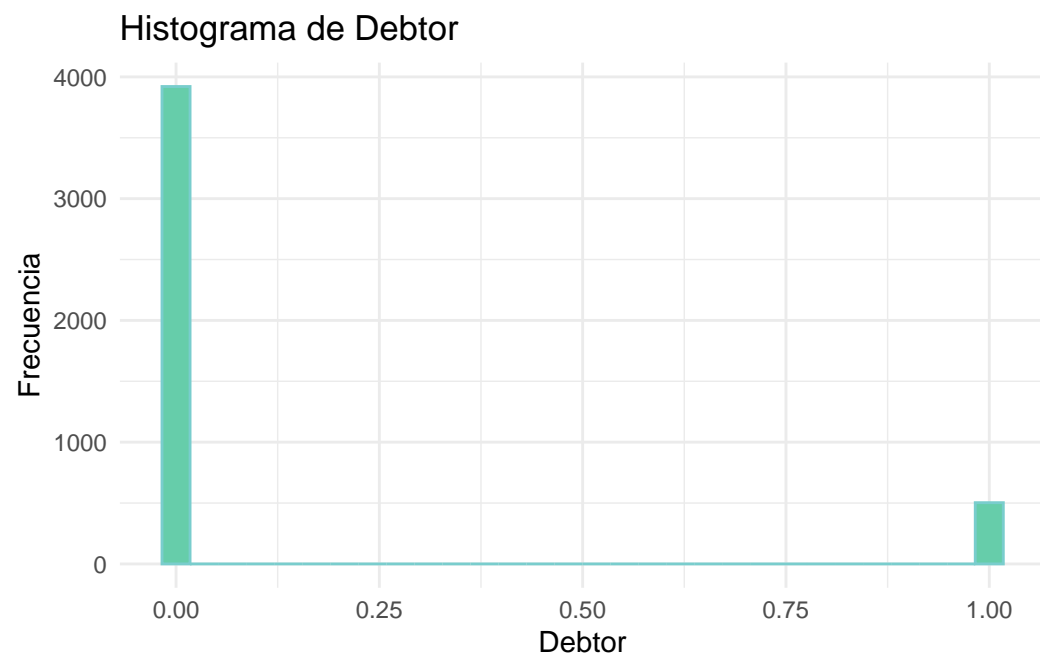
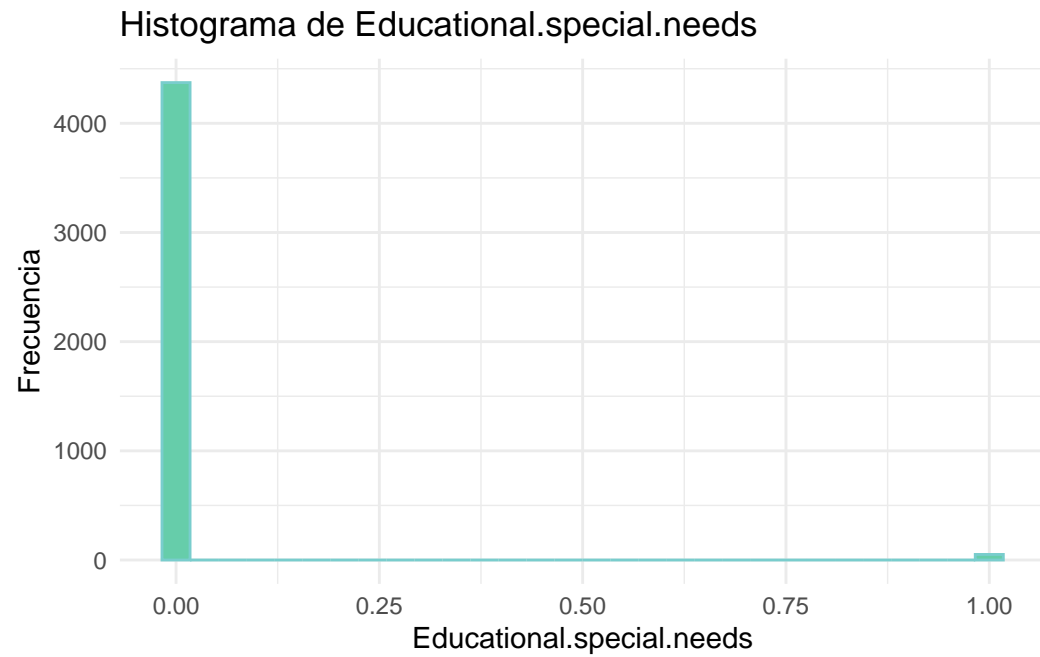
Histograma de Previous.qualification

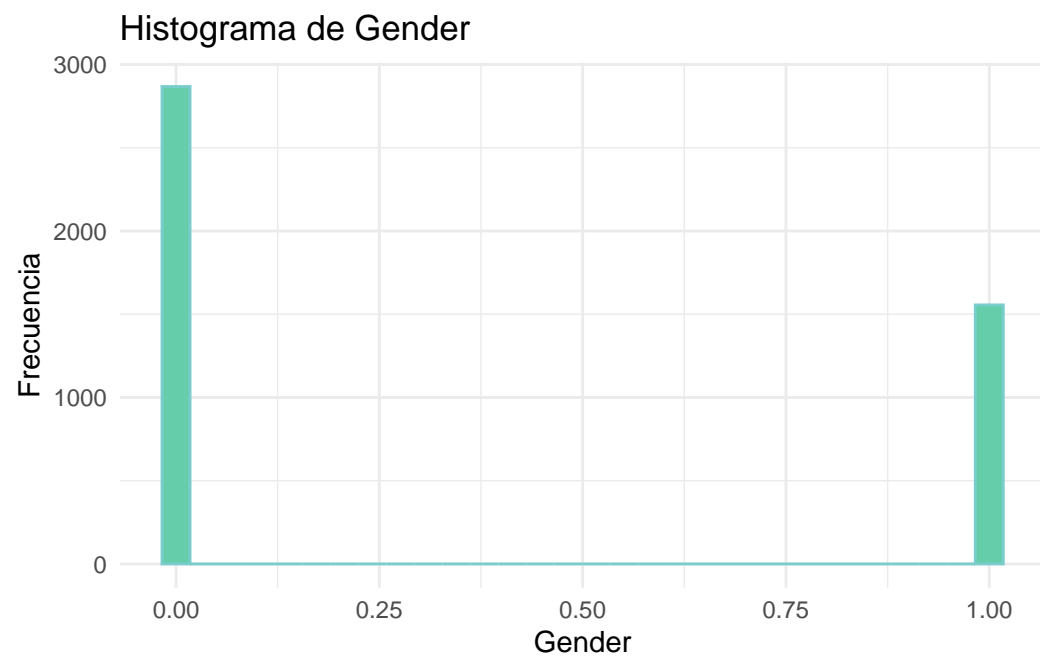
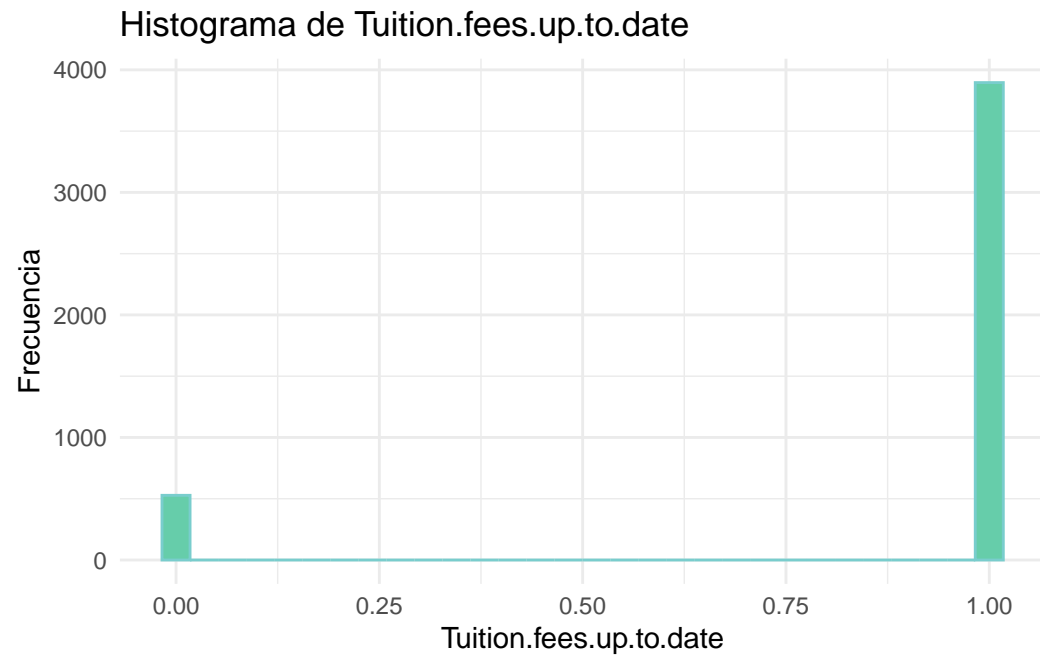




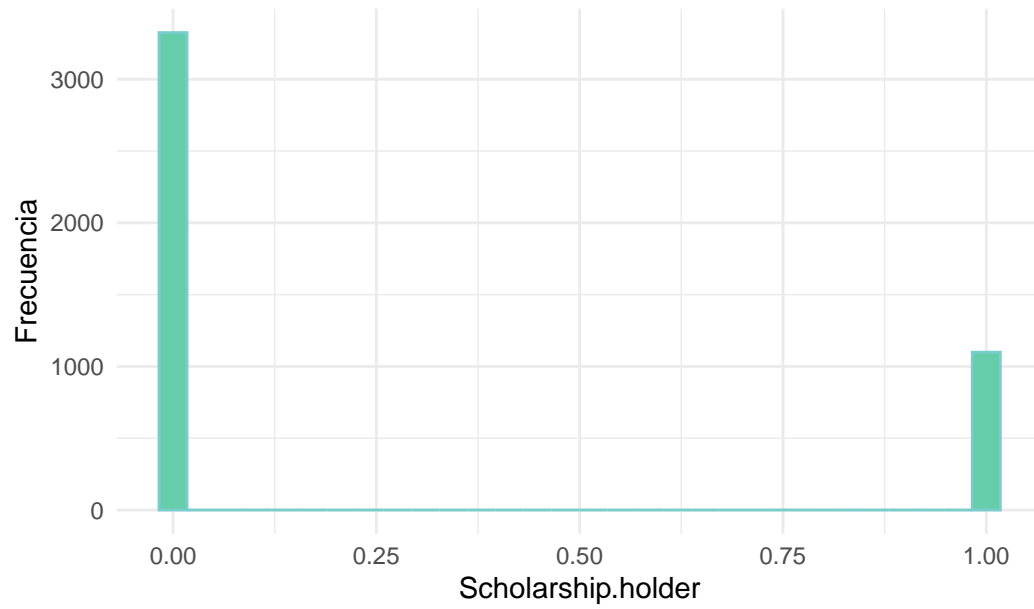




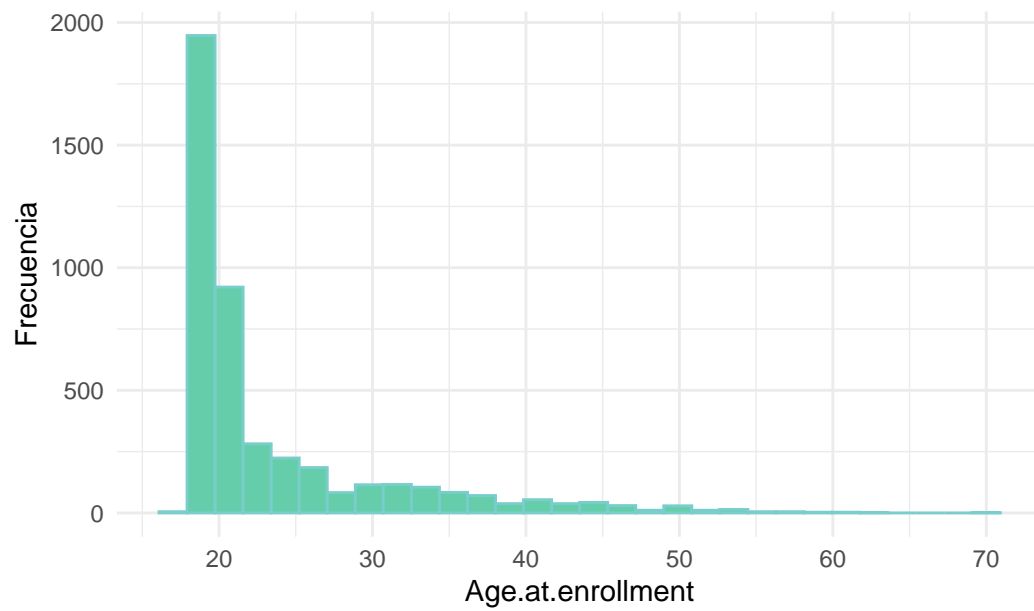


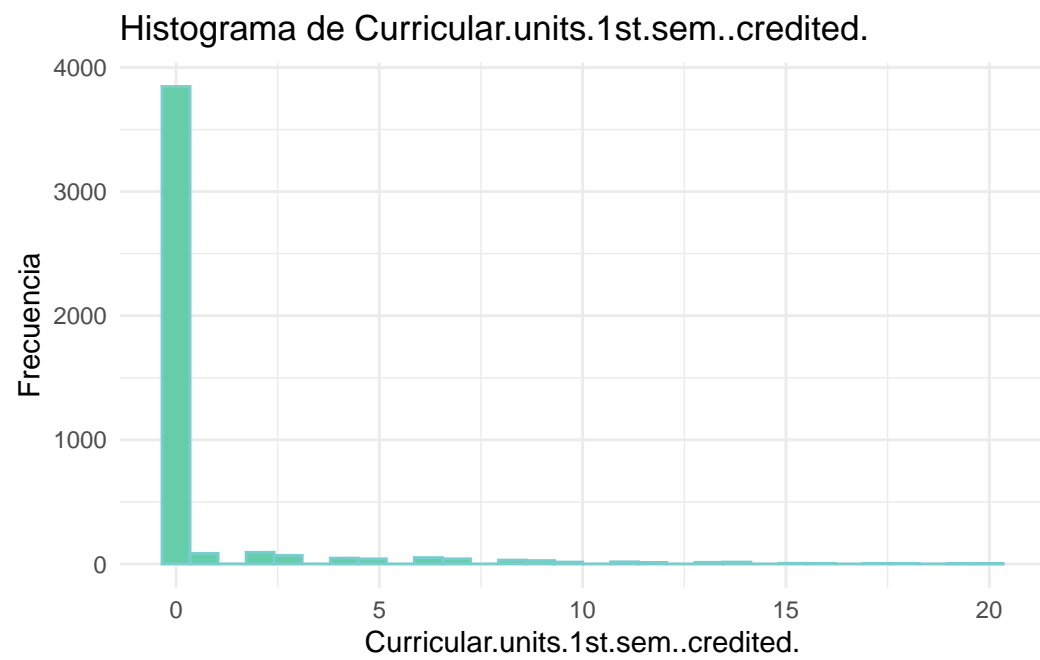
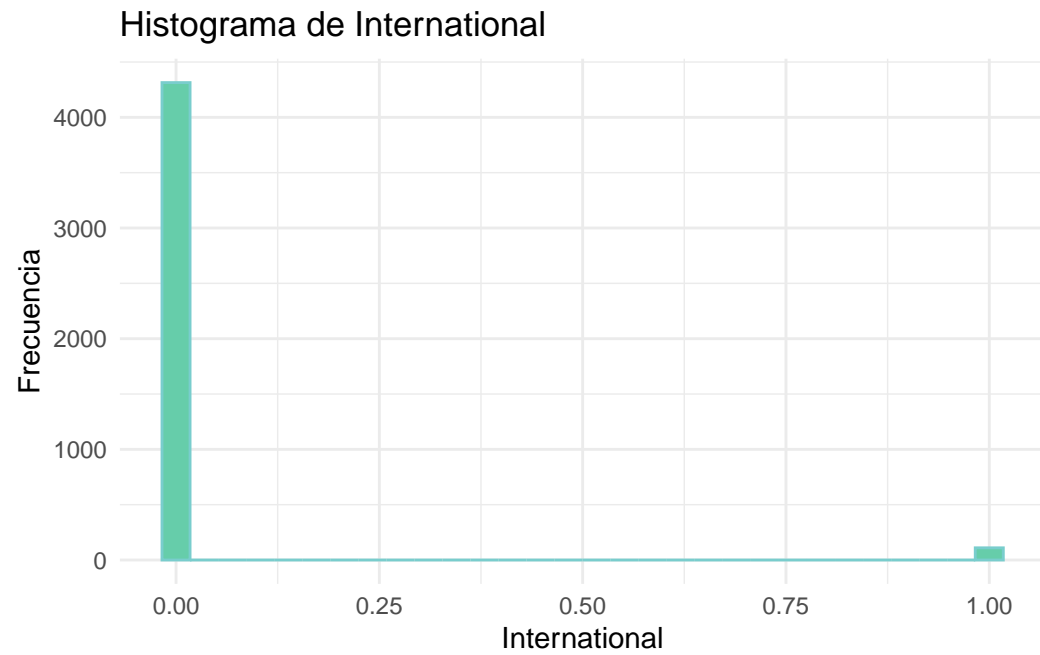


Histograma de Scholarship.holder

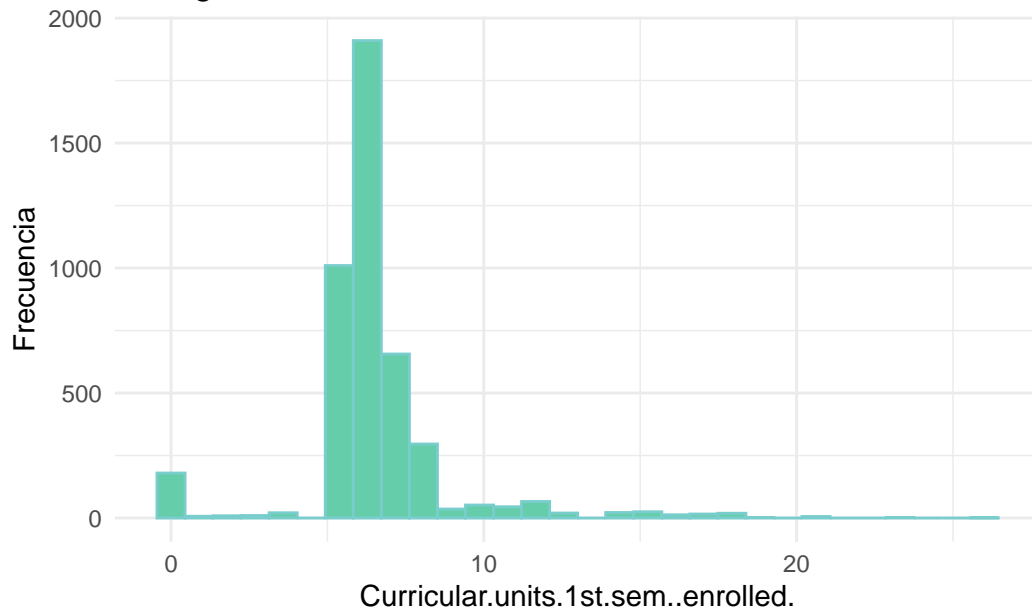


Histograma de Age.at.enrollment

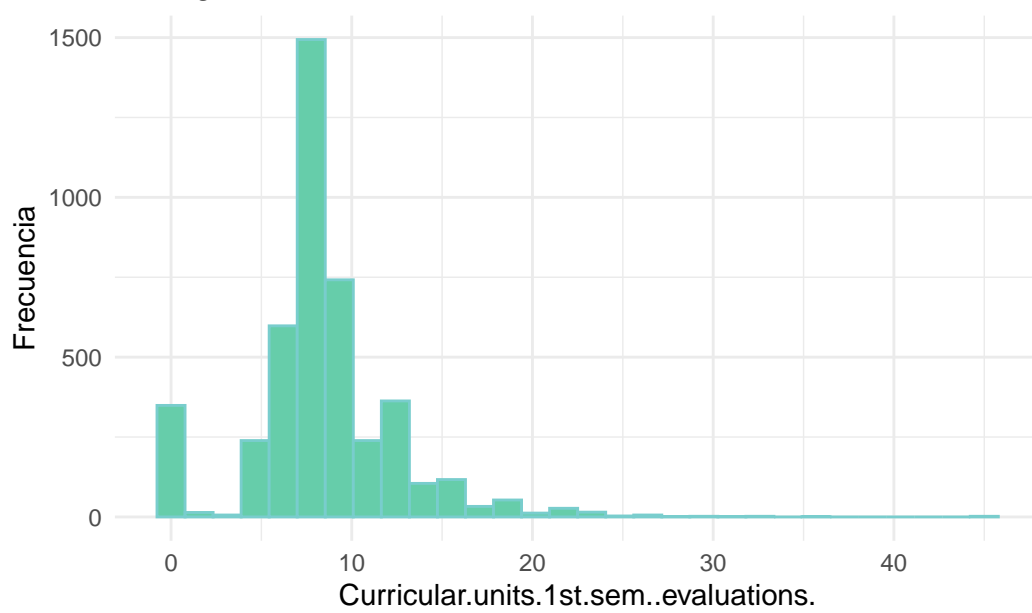




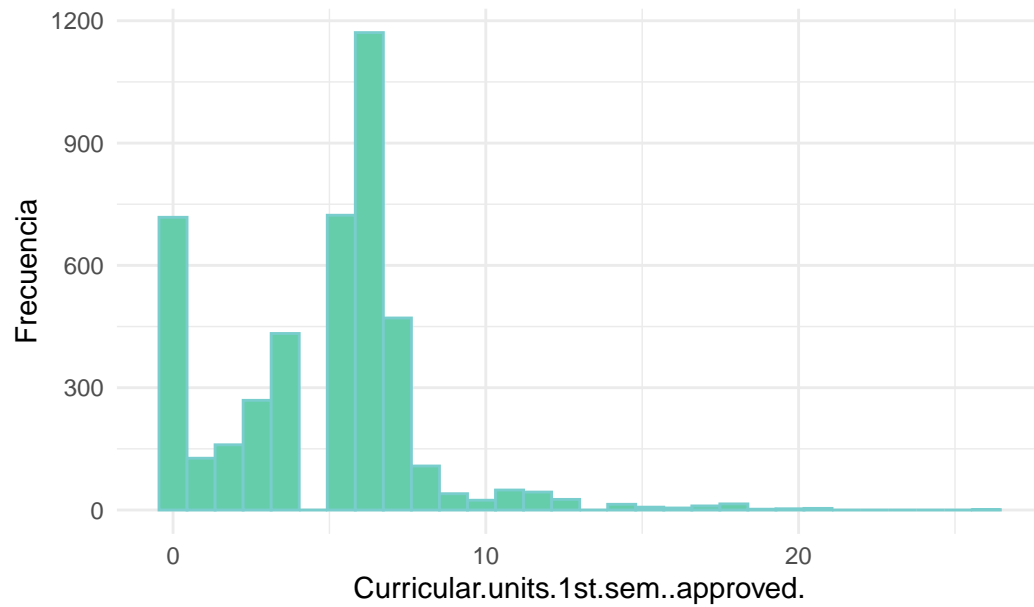
Histograma de Curricular.units.1st.sem..enrolled.



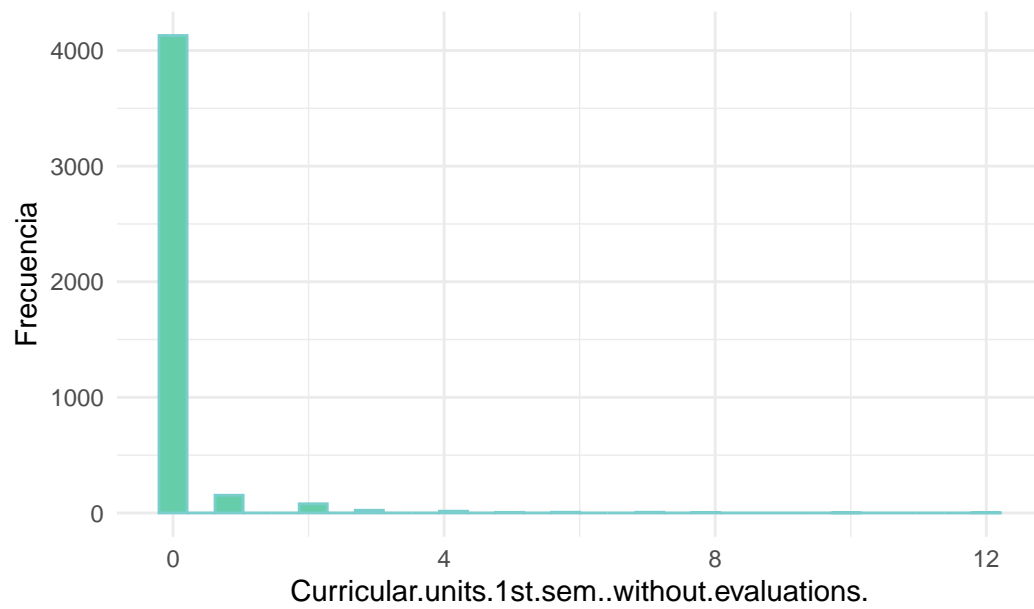
Histograma de Curricular.units.1st.sem..evaluations.



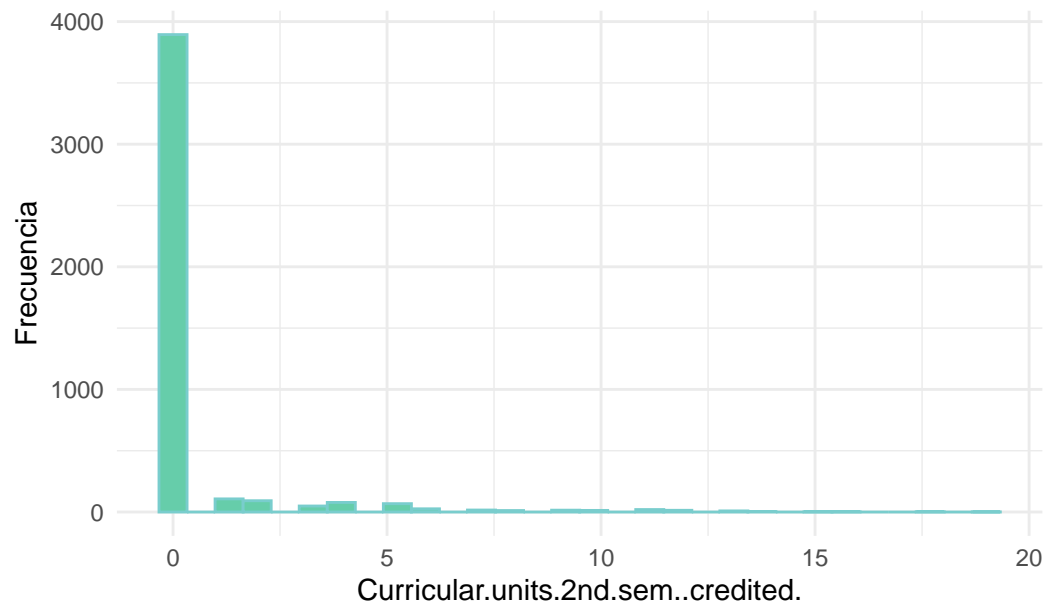
Histograma de Curricular.units.1st.sem..approved.



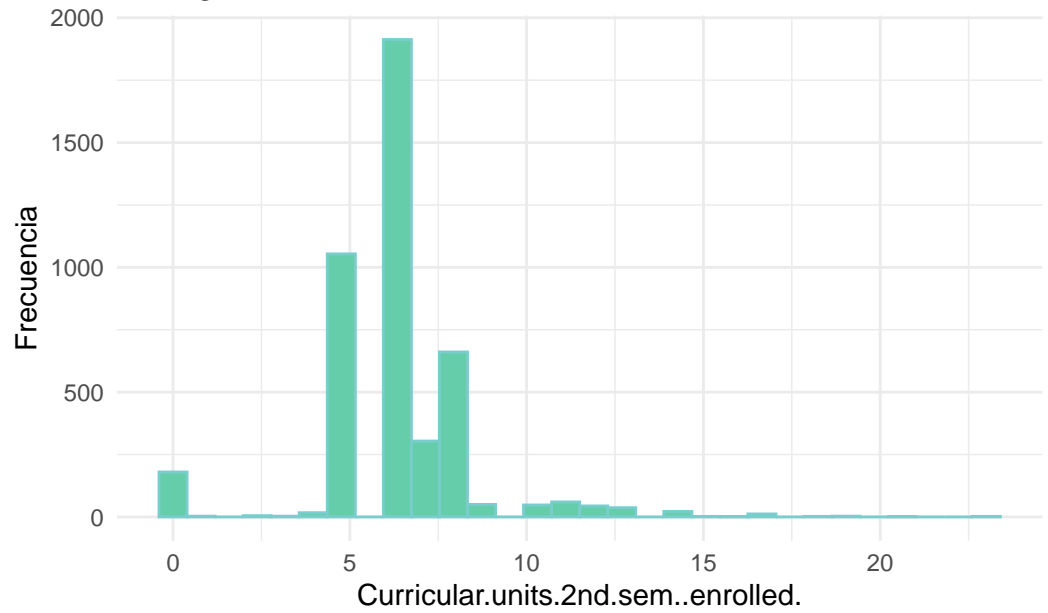
Histograma de Curricular.units.1st.sem..without.evaluations.



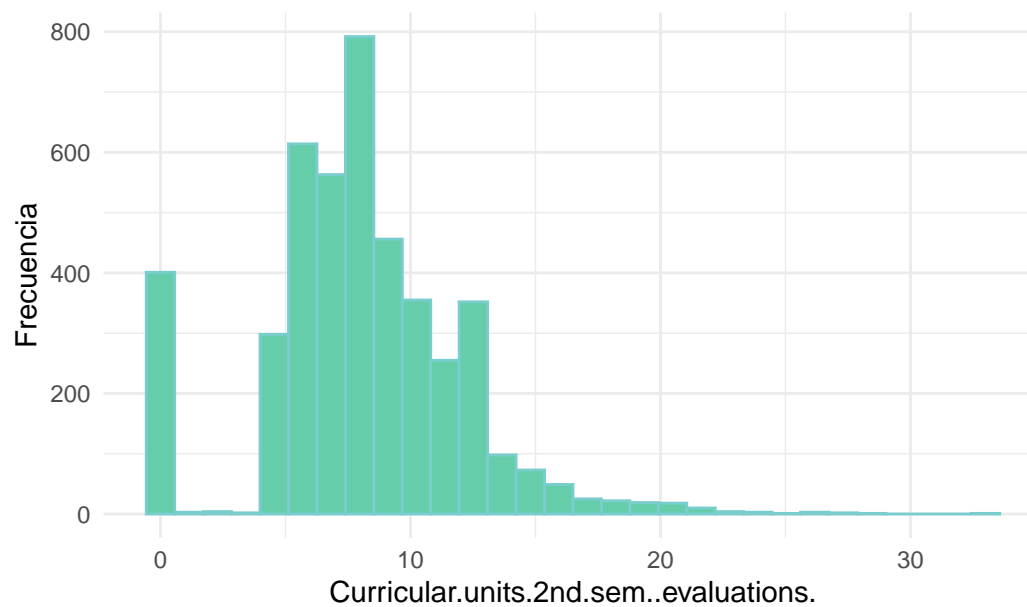
Histograma de Curricular.units.2nd.sem..credited.



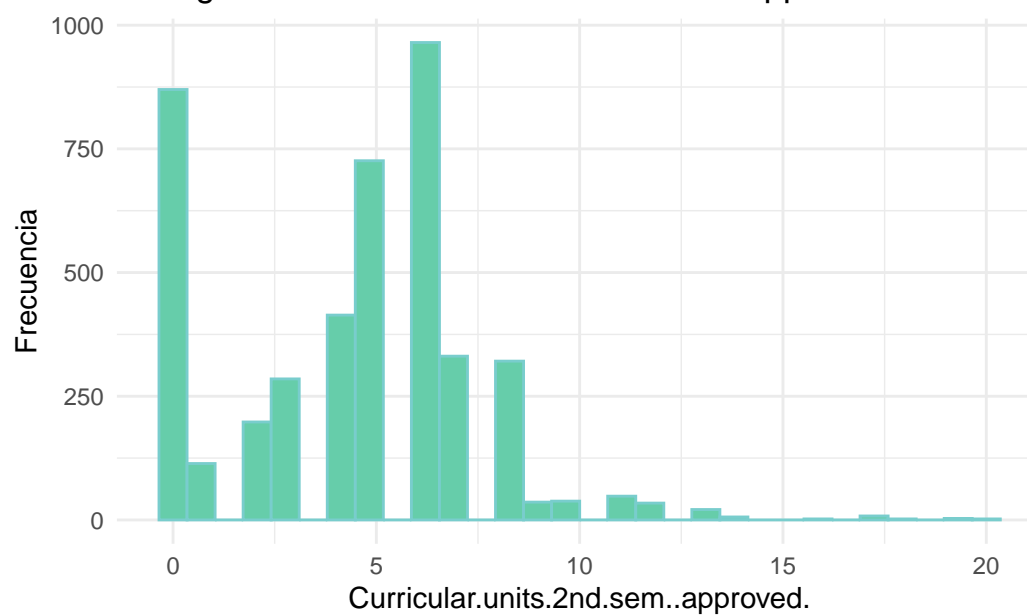
Histograma de Curricular.units.2nd.sem..enrolled.

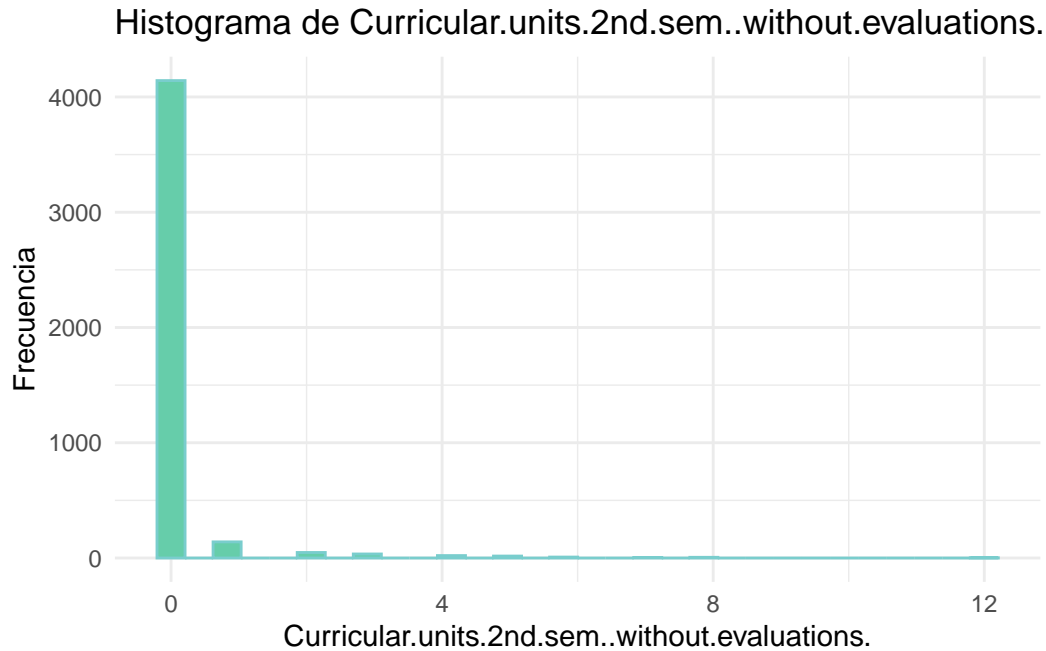


Histograma de Curricular.units.2nd.sem..evaluations.



Histograma de Curricular.units.2nd.sem..approved.





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

```
grafico_internacional_genero <- function(df, genero, internacional, pais) {
  df_plot <- df %>%
    mutate(
      genero = factor({{genero}}, levels = c(1, 0), labels = c("Hombres", "Mujeres")),
      tipo = factor({{internacional}}, levels = c(0, 1), labels = c("Nacional", "Internacional"))
    ) %>%
    group_by(genero, tipo) %>%
    summarise(Frecuencia = n(), .groups = "drop")

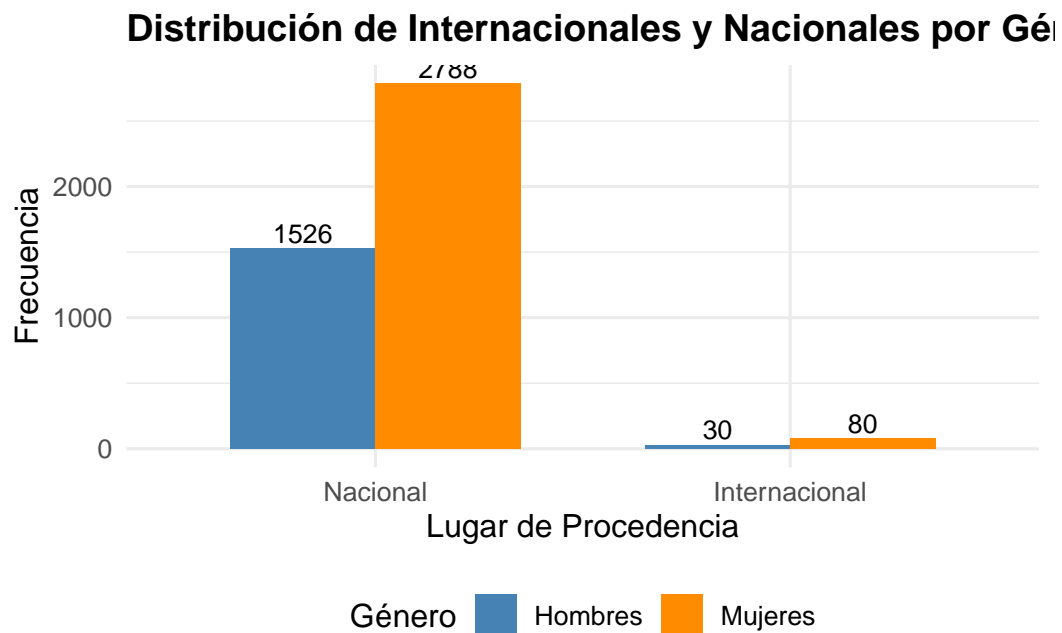
  ggplot(df_plot, aes(x = tipo, y = Frecuencia, fill = genero)) +
    geom_col(position = "dodge", width = 0.7) +
    geom_text(aes(label = Frecuencia),
              position = position_dodge(width = 0.7),
              vjust = -0.3, size = 3.5) +
    scale_fill_manual(values = c("steelblue", "darkorange")) +
    labs(
```

```

    title = "Distribución de Internacionales y Nacionales por Género",
    x = "Lugar de Procedencia",
    y = "Frecuencia",
    fill = "Género"
  ) +
  theme_minimal(base_size = 12) +
  theme(
    legend.position = "bottom",
    plot.title = element_text(face = "bold"),
    axis.text.x = element_text(angle = 0, hjust = 0.5)
  )
}

grafico_internacional_genero(datos, datos$Gender, datos$International, datos$Nationality

```



```

gráfico_necesidad_genero <- function(df, genero, needs){
  df_plot <- df %>%
    mutate(
      genero = factor({{genero}}, levels = c(1, 0), labels = c("Hombres", "Mujeres")),

```

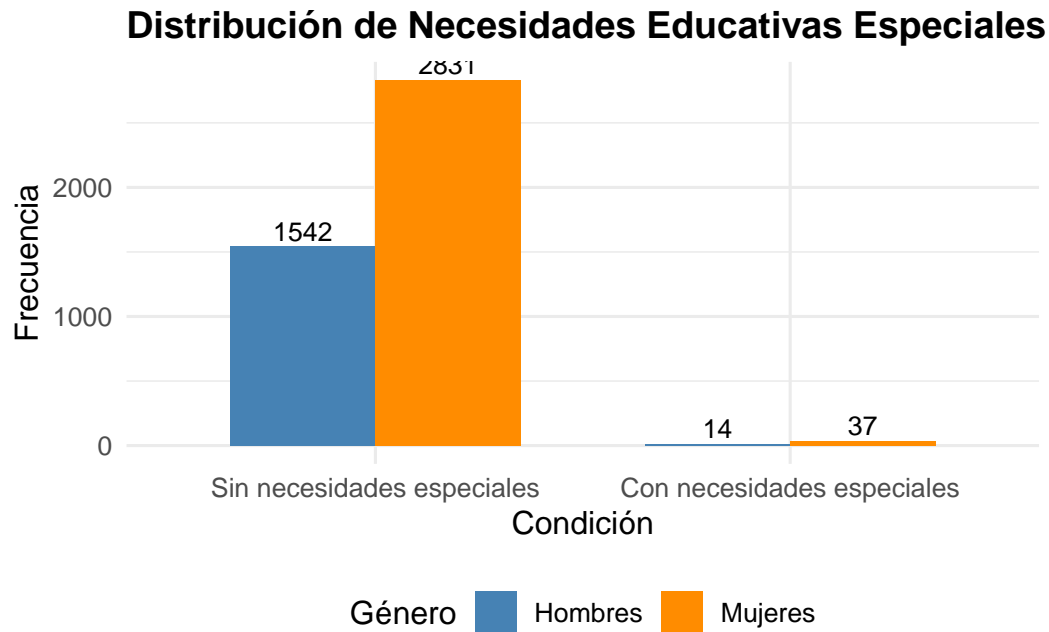
```

    necesidades = factor({{needs}}, levels = c(0, 1),
    labels = c("Sin necesidades especiales", "Con necesidades especiales"))
  ) %>%
  group_by(genero, necesidades) %>%
  summarise(Frecuencia = n(), .groups = "drop")

ggplot(df_plot, aes(x = necesidades, y = Frecuencia, fill = genero)) +
  geom_col(position = "dodge", width = 0.7) +
  geom_text(aes(label = Frecuencia),
  position = position_dodge(width = 0.7),
  vjust = -0.3, size = 3.5) +
  scale_fill_manual(values = c("steelblue", "darkorange")) +
  labs(
    title = "Distribución de Necesidades Educativas Especiales por Género",
    x = "Condición",
    y = "Frecuencia",
    fill = "Género"
  ) +
  theme_minimal(base_size = 12) +
  theme(
    legend.position = "bottom",
    plot.title = element_text(face = "bold"),
    axis.text.x = element_text(angle = 0, hjust = 0.5)
  )
}

gráfico_necesidad_genero(datos, datos$Gender, datos$Educational.special.needs)

```

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

7- Identificar valores faltantes y posibles outliers.

```
datos_faltantes <- datos %>%
  filter(if_any(everything(), is.na))
head(datos_faltantes)
```

```
[1] Marital.status
[2] Application.mode
[3] Application.order
[4] Course
[5] Daytime.evening.attendance.
[6] Previous.qualification
[7] Previous.qualification..grade.
[8] Nationality
[9] Mother.s.qualification
[10] Father.s.qualification
[11] Mother.s.occupation
```

```

[12] Father.s.occupation
[13] Admission.grade
[14] Displaced
[15] Educational.special.needs
[16] Debtor
[17] Tuition.fees.up.to.date
[18] Gender
[19] Scholarship.holder
[20] Age.at.enrollment
[21] International
[22] Curricular.units.1st.sem..credited.
[23] Curricular.units.1st.sem..enrolled.
[24] Curricular.units.1st.sem..evaluations.
[25] Curricular.units.1st.sem..approved.
[26] Curricular.units.1st.sem..grade.
[27] Curricular.units.1st.sem..without.evaluations.
[28] Curricular.units.2nd.sem..credited.
[29] Curricular.units.2nd.sem..enrolled.
[30] Curricular.units.2nd.sem..evaluations.
[31] Curricular.units.2nd.sem..approved.
[32] Curricular.units.2nd.sem..grade.
[33] Curricular.units.2nd.sem..without.evaluations.
[34] Unemployment.rate
[35] Inflation.rate
[36] GDP
[37] Target
<0 rows> (o 0- extensión row.names)

```

```

datos %>%
  summarise(

```

```

across(
  where(is.numeric),
  ~sum(
    .<quantile(.,0.25,na.rm=TRUE)-1.5*IQR(.)|
    .>quantile(.,0.75,na.rm=TRUE)+1.5*IQR(.),na.rm = TRUE
  )
)
)#cantidad de outliers por variable

```

	Marital.status	Application.mode	Application.order	Course
1	505	0	541	442
	Daytime.evening.attendance.	Previous.qualification	Nacionality	
1	483	707	110	
	Mother.s.qualification	Father.s.qualification	Mother.s.occupation	
1	0	0	182	
	Father.s.occupation	Displaced	Educational.special.needs	Debtor
1	177	0	51	503
	Tuition.fees.up.to.date	Gender	Scholarship.holder	Age.at.enrollment
1	528	0	1099	441
	International	Curricular.units.1st.sem..credited.		
1	110	577		
	Curricular.units.1st.sem..enrolled.	Curricular.units.1st.sem..evaluations.		
1	424	158		
	Curricular.units.1st.sem..approved.			
1	180			
	Curricular.units.1st.sem..without.evaluations.			
1	294			
	Curricular.units.2nd.sem..credited.	Curricular.units.2nd.sem..enrolled.		
1	530	369		

	Curricular.units.2nd.sem..evaluations.	Curricular.units.2nd.sem..approved.
1	109	44

	Curricular.units.2nd.sem..without.evaluations.
1	282

```
es_outlier <- function(x) {

  if(!is.numeric(x)) return(rep(FALSE,length(x)))
  q1 <- quantile(x,0.25,na.rm = TRUE)
  q3 <- quantile(x,0.75,na.rm = TRUE)

  resultado <- x<(q1-1.5*IQR(x,na.rm = TRUE)) |
               x>(q3+1.5*IQR(x,na.rm = TRUE))
  return(resultado)
}

outliers <- as.data.frame(sapply(datos,es_outlier))
head(outliers)
```

	Marital.status	Application.mode	Application.order	Course
1	FALSE	FALSE	TRUE	TRUE
2	FALSE	FALSE	FALSE	FALSE
3	FALSE	FALSE	TRUE	FALSE
4	FALSE	FALSE	FALSE	FALSE
5	TRUE	FALSE	FALSE	TRUE
6	TRUE	FALSE	FALSE	FALSE

	Daytime.evening.attendance.	Previous.qualification
1	FALSE	FALSE
2	FALSE	FALSE
3	FALSE	FALSE

4	FALSE	FALSE
5	TRUE	FALSE
6	TRUE	TRUE

	Previous.qualification..grade.	Nacionality	Mother.s.qualification
1	FALSE	FALSE	FALSE
2	FALSE	FALSE	FALSE
3	FALSE	FALSE	FALSE
4	FALSE	FALSE	FALSE
5	FALSE	FALSE	FALSE
6	FALSE	FALSE	FALSE

	Father.s.qualification	Mother.s.occupation	Father.s.occupation
1	FALSE	FALSE	FALSE
2	FALSE	FALSE	FALSE
3	FALSE	FALSE	FALSE
4	FALSE	FALSE	FALSE
5	FALSE	FALSE	FALSE
6	FALSE	FALSE	FALSE

	Admission.grade	Displaced	Educational.special.needs	Debtor
1	FALSE	FALSE	FALSE	FALSE
2	FALSE	FALSE	FALSE	FALSE
3	FALSE	FALSE	FALSE	FALSE
4	FALSE	FALSE	FALSE	FALSE
5	FALSE	FALSE	FALSE	FALSE
6	FALSE	FALSE	FALSE	TRUE

	Tuition.fees.up.to.date	Gender	Scholarship.holder	Age.at.enrollment
1	FALSE	FALSE	FALSE	FALSE
2	TRUE	FALSE	FALSE	FALSE
3	TRUE	FALSE	FALSE	FALSE
4	FALSE	FALSE	FALSE	FALSE
5	FALSE	FALSE	FALSE	TRUE

6	FALSE	FALSE	FALSE	TRUE
International Curricular.units.1st.sem..credited.				
1	FALSE		FALSE	
2	FALSE		FALSE	
3	FALSE		FALSE	
4	FALSE		FALSE	
5	FALSE		FALSE	
6	FALSE		FALSE	
Curricular.units.1st.sem..enrolled. Curricular.units.1st.sem..evaluations.				
1		TRUE		FALSE
2		FALSE		FALSE
3		FALSE		FALSE
4		FALSE		FALSE
5		FALSE		FALSE
6		FALSE		FALSE
Curricular.units.1st.sem..approved. Curricular.units.1st.sem..grade.				
1		FALSE		FALSE
2		FALSE		FALSE
3		FALSE		FALSE
4		FALSE		FALSE
5		FALSE		FALSE
6		FALSE		FALSE
Curricular.units.1st.sem..without.evaluations.				
1			FALSE	
2			FALSE	
3			FALSE	
4			FALSE	
5			FALSE	
6			FALSE	
Curricular.units.2nd.sem..credited. Curricular.units.2nd.sem..enrolled.				

1	FALSE	TRUE
2	FALSE	FALSE
3	FALSE	FALSE
4	FALSE	FALSE
5	FALSE	FALSE
6	FALSE	FALSE

Curricular.units.2nd.sem..evaluations. Curricular.units.2nd.sem..approved.

1	FALSE	FALSE
2	FALSE	FALSE
3	FALSE	FALSE
4	FALSE	FALSE
5	FALSE	FALSE
6	TRUE	FALSE

Curricular.units.2nd.sem..grade.

1	FALSE
2	FALSE
3	FALSE
4	FALSE
5	FALSE
6	FALSE

Curricular.units.2nd.sem..without.evaluations. Unemployment.rate

1	FALSE	FALSE
2	FALSE	FALSE
3	FALSE	FALSE
4	FALSE	FALSE
5	FALSE	FALSE
6	TRUE	FALSE

Inflation.rate GDP Target

1	FALSE	FALSE	FALSE
2	FALSE	FALSE	FALSE

3	FALSE	FALSE	FALSE
4	FALSE	FALSE	FALSE
5	FALSE	FALSE	FALSE
6	FALSE	FALSE	FALSE

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

9- Bibliografía

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

https://rpubs.com/Elyn1017/Aunivariado_Vcuantitativas_CasoMedicos

<https://www.uca.edu.sv/mpe/wp-content/uploads/2020/09/61.-Hernandez-W-y-Montano-Y.-2020-Analisis-de-la-desercion-escolar.pdf>