

ACTIVIDAD 2

GRUPO 6



```
---
```

```
title: "Cancer Cervical"
```

```
format: html
```

```
editor: visual
```

```
---
```

GRUPO 06

Choquecahua Bendezú Carol Neyduth

Clemente Valenzuela Brithney Coraima

Cortez Carbonell Dariana Ysabel

Felix Yataco Maria Fernanda

Huaripuma Lozano Anyelina Yuli

Larico Mamani Liz Heydi Patricia



UNIVERSIDAD PRIVADA
SAN JUAN BAUTISTA

Instalar paquetes

```
{r}
install.packages("tidyverse")
install.packages("rio")
install.packages("here")
install.packages("janitor")
install.packages("skimr")
install.packages("visdat")
install.packages("DataExplorer")
```

Cargar paquetes

```
{r}
library(rio)
library(here)
library(janitor)
library(skimr)
library(visdat)
library(DataExplorer)
library(ggplot2)
```

CAPTURAS DEL TRABAJO

The screenshot shows the RStudio interface with two tabs open: "ACTIVIDAD 2 conoc_actit_factor_cancer..." and "ACTIVIDAD_2__conoc_actit_factor_canc...". The left pane displays R code:

```
{r}
conoc_actit_factor_cancer_cervical = import(here("data", "conoc_actit_factor_cancer_cervical.csv"))
```

Below the code, a section titled "Vistazo al contenido" (Preview of content) contains the following text:

¿Cuántas variables y observaciones hay?

El primer número indica el número de filas, el segundo, el número de columnas.

```
{r}
dim(conoc_actit_factor_cancer_cervical)
```

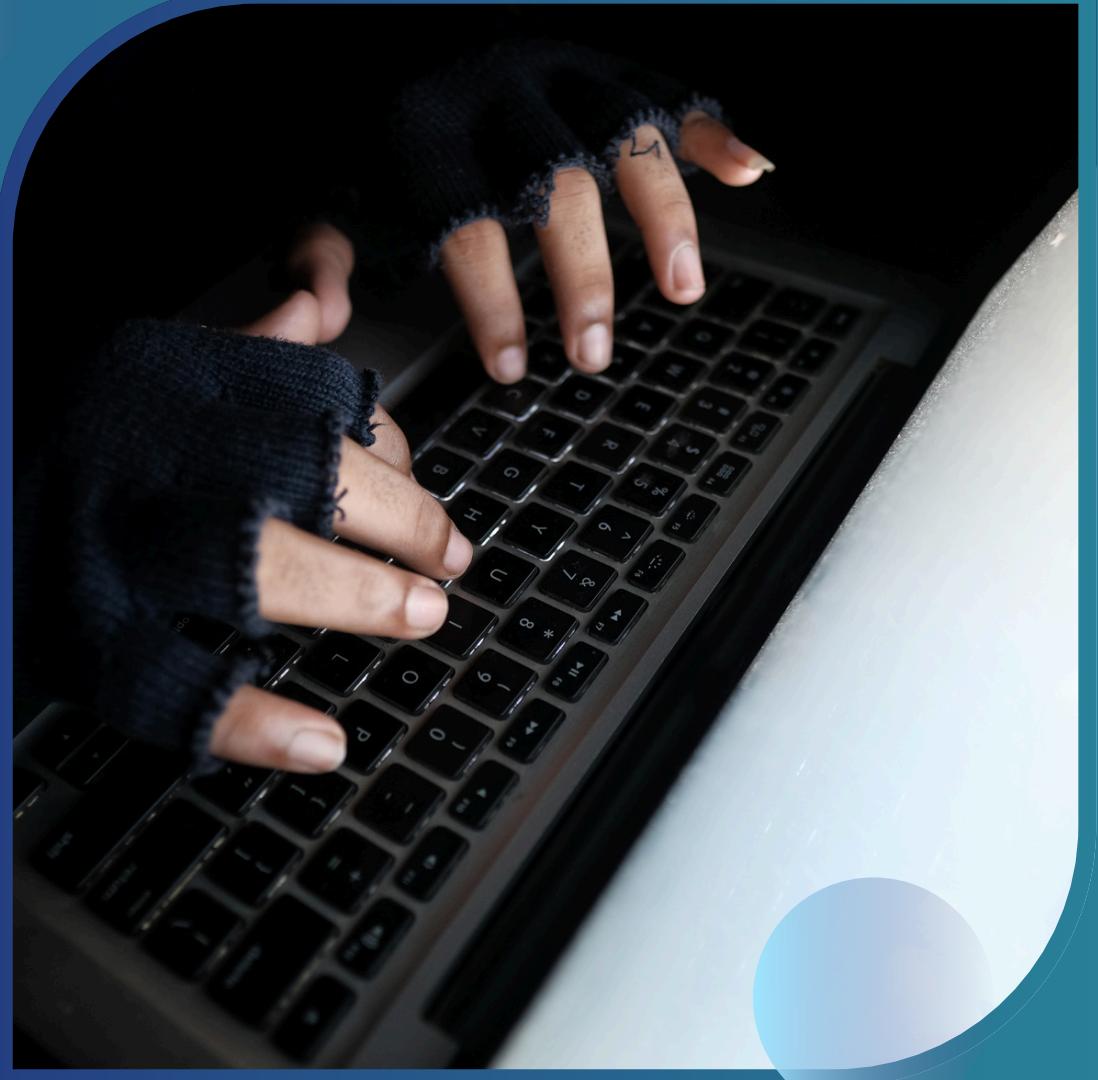
Below this, another section titled "¿Cuántas y qué tipos de variables hay?" (How many and what types of variables are there?) contains the following text:

```
{r}
str(conoc_actit_factor_cancer_cervical)
```

The output of the str() command shows the structure of the data frame:

```
'data.frame': 218 obs. of 18 variables:
 $ paciente_num    : int 1 2 3 4 5 6 7 8 9 10 ...
 $ edad            : int 53 54 26 25 25 38 29 39 50 25 ...
 $ e_marital       : chr "casada" "casada" "soltera" "soltera" ...
 $ n_educacion     : chr "superior" "superior" "superior" "superior" ...
 $ religion         : chr "catolico" "catolico" "catolico" "ninguna" ...
 $ etnia            : chr "mestizo" "mestizo" "blanco" "mestizo" ...
 $ procedencia      : chr "urbano" "rural" "urbano" "rural" ...
 $ ocupacion        : chr "otro" "empleada" "estudiante" "sin empleo" ...
 $ ocupacion_convi  : chr "empleado" "empleado" "estudiante" "sin empleo" ...
 $ antec_fam        : chr "no" "no" "no" "no" ...
 $ edad_relacion_sexual: int 23 18 18 18 18 20 16 17 19 17 ...
 $ parejas_sex      : int 2 1 4 10 6 2 3 4 3 2 ...
```

At the bottom left of the RStudio interface, there is a small note: "# Visualmente" followed by a dropdown arrow.



RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

+ Go to file/function Addins

Project: (None)

ACTIVIDAD 2 conoc_actit_factor_cancer... ACTIVIDAD_2_conoc_actit_factor_canc...
Render on Save ABC Render  Run Publish Outline

Source Visual B I Header 3 Format Insert Table

```
$ actitud : chr "negativa" "positiva" "positiva" "positiva" ...
$ practica : chr "incorrecta" "incorrecta" "incorrecta" "correcta" ...
```

Una función similar

```
{r}
dplyr::glimpse(conoc_actit_factor_cancer_cervical)

Rows: 218
Columns: 18
$ paciente_num <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, ...
$ edad <int> 53, 54, 26, 25, 25, 38, 29, 39, 50, 25, 26, 28, 25, 27, 28, 32, ...
$ e_marital <chr> "casada", "casada", "soltera", "soltera", "soltera", ...
$ n_educacion <chr> "superior", "superior", "superior", "superior", "superior", "su...
$ religion <chr> "catolico", "catolico", "catolico", "ninguna", "ninguna", "cato...
$ etnia <chr> "mestizo", "mestizo", "blanco", "mestizo", "mestizo", ...
$ procedencia <chr> "urbano", "rural", "urbano", "rural", "rural", "urbano...
$ ocupacion <chr> "otro", "empleada", "estudiante", "sin empleo", "estudiante", ...
$ ocupacion_convi <chr> "empleado", "empleado", "estudiante", "sin empleo", "estudiante...
$ antec_fam <chr> "no", "no", "no", "no", "no", "si", "si", "no", "no", ...
$ edad_relacion_sexual <int> 23, 18, 18, 18, 18, 20, 16, 17, 19, 17, 17, 26, 20, 15, 17, 18, ...
$ parejas_sex <int> 2, 1, 4, 10, 6, 2, 3, 4, 3, 2, 5, 1, 2, 2, 2, 5, 1, 1, 2, 1, 2, ...
$ num_hijos <int> 3, 3, 2, 0, 0, 2, 2, 4, 2, 2, 1, 0, 0, 0, 4, 1, 1, 1, 0, 0, 2, ...
$ met_anticoncep <chr> "no uso", "no uso", "no uso", "no uso", "no uso", "no uso", ...
$ antec_ets <chr> "no", "no", "si", "no", "no", "no", "no", "no", "no", "no", ...
$ conocimiento <chr> "alto", "medio", "medio", "medio", "medio", "alto", "alto", "ba...
$ actitud <chr> "negativa", "positiva", "positiva", "positiva", "positiva", "ne...
$ practica <chr> "incorrecta", "incorrecta", "incorrecta", "correcta", "correcta", ...
```

Limpieza de ...
Paso uno: ...
Paso dos: c...
Paso tres: ...
Optimizand...
Corregir n...
Paso 4: cor...
Inspeció...
Paso 5: Col...
Paso 6: Tra...
Paso 7: Re...

Estadísticos descriptivos y otros parámetros para exploración de datos

```
{r}
skimr::skim(conoc_actit_factor_cancer_cervical)
```

Visualmente Quarto

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

ACTIVIDAD 2 conoc_actit_factor_cancer... ACTIVIDAD_2_conoc_actit_factor_canc...

Source Visual B I </> Header 3 | Format | Insert | Table

```
$ conocimiento      <chr> "alto", "medio", "medio", "medio", "medio", "alto", "alto", "ba...
$ actitud          <chr> "negativa", "positiva", "positiva", "positiva", "positiva", "ne...
$ practica         <chr> "incorrecta", "incorrecta", "incorrecta", "correcta", "correcta...
```

Estadísticos descriptivos y otros parámetros para exploración de datos

```
{r}
skimr::skim(conoc_actit_factor_cancer_cervical)
```

A tibble: 5 × 11

	skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	▸
1	paciente_num	0	1	109.500000	63.075352	1	55.25	109.5	163.75	
2	edad	0	1	45.408257	11.302433	25	38.25	48.0	54.00	
3	edad_relacion...	0	1	19.325688	2.321882	13	18.00	19.0	20.00	
4	parejas_sex	0	1	2.509174	1.699121	0	2.00	2.0	3.00	
5	num_hijos	0	1	2.059633	1.323702	0	1.00	2.0	3.00	

5 rows | 1-10 of 11 columns

Resumen por variable

```
{r}
summary(conoc_actit_factor_cancer_cervical)
```

paciente_num edad c_mortal n_eduacion relacion

Visualmente ▾

Console

Project: (None) ▾

Run ▾ Publish ▾ Outline ▾

Instalar paq... Cargar paq... Importando... Vistazo al c... ¿Cuántas ... ¿Cuantas ... Visualmente Limpieza de ... Paso uno: ... Paso dos: c... Paso tres: ... Optimizand... Corregir n... Paso 4: cor... Inspecció... Paso 5: Col... Paso 6: Tra... Paso 7: Re...

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

+ Go to file/function Addins Project: (None)

ACTIVIDAD 2 conoc_actit_factor_cancer... ACTIVIDAD_2__conoc_actit_factor_canc...

Source Visual Header 3 Insert Table

Resumen por variable

```
{r}
summary(conoc_actit_factor_cancer_cervical)
```

paciente_num	edad	e_marital	n_educacion	religion
Min. : 1.00	Min. :25.00	Length:218	Length:218	Length:218
1st Qu.: 55.25	1st Qu.:38.25	Class :character	Class :character	Class :character
Median :109.50	Median :48.00	Mode :character	Mode :character	Mode :character
Mean :109.50	Mean :45.41			
3rd Qu.:163.75	3rd Qu.:54.00			
Max. :218.00	Max. :67.00			

etnia	procedencia	ocupacion	ocupacion_convi
Length:218	Length:218	Length:218	Length:218
Class :character	Class :character	Class :character	Class :character
Mode :character	Mode :character	Mode :character	Mode :character

antec_fam	edad_relacion_sexual	parejas_sex	num_hijos	met_anticoncep
Length:218	Min. :13.00	Min. : 0.000	Min. :0.00	Length:218
Class :character	1st Qu.:18.00	1st Qu.: 2.000	1st Qu.:1.00	Class :character
Mode :character	Median :19.00	Median : 2.000	Median :2.00	Mode :character
	Mean :19.33	Mean : 2.509	Mean :2.06	
	3rd Qu.:20.00	3rd Qu.: 3.000	3rd Qu.:3.00	
	Max. :30.00	Max. :15.000	Max. :6.00	

antec_ets	conocimiento	actitud	practica
Length:218	Length:218	Length:218	Length:218
Class :character	Class :character	Class :character	Class :character
Mode :character	Mode :character	Mode :character	Mode :character

Instalar paq...
Cargar paq...
Importando...
Vistazo al c...
¿Cuántas ...
¿Cuantas ...
Visualmente
Limpieza de ...
Paso uno: ...
Paso dos: c...
Paso tres: ...
Optimizand...
Corregir n...
Paso 4: cor...
Inspeció...
Paso 5: Col...
Paso 6: Tra...
Paso 7: Re...

Visualmente Quarto

Console

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

+ Go to file/function Addins

ACTIVIDAD 2 conoc_actit_factor_cancer... ACTIVIDAD_2__conoc_actit_factor_canc...

Render on Save ABC Render Header 3 Format Insert Table

Source Visual B I </> Header 3 Format Insert Table

Visualmente

```
{r}
View(conoc_actit_factor_cancer_cervical)
```

Error in View : objeto 'conoc_actit_factor_cancer_cervical' no encontrado

```
{r}
visdat::vis_dat(conoc_actit_factor_cancer_cervical)
```



Observations

Type

character

integer

Visualmente Quarto

Console

Project: (None)

Run Publish Outline

Instalar paq... Cargar paq... Importando... Vistazo al c... ¿Cuántas ... ¿Cuantas ... Visualmente Limpieza de ... Paso uno: ... Paso dos: c... Paso tres: ... Optimizand... Corregir n... Paso 4: cor... Inspecció... Paso 5: Col... Paso 6: Tra... Paso 7: Re...

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

ACTIVIDAD 2 conoc_actit_factor_cancer... ACTIVIDAD_2__conoc_actit_factor_canc...

Render on Save ABC Render Run

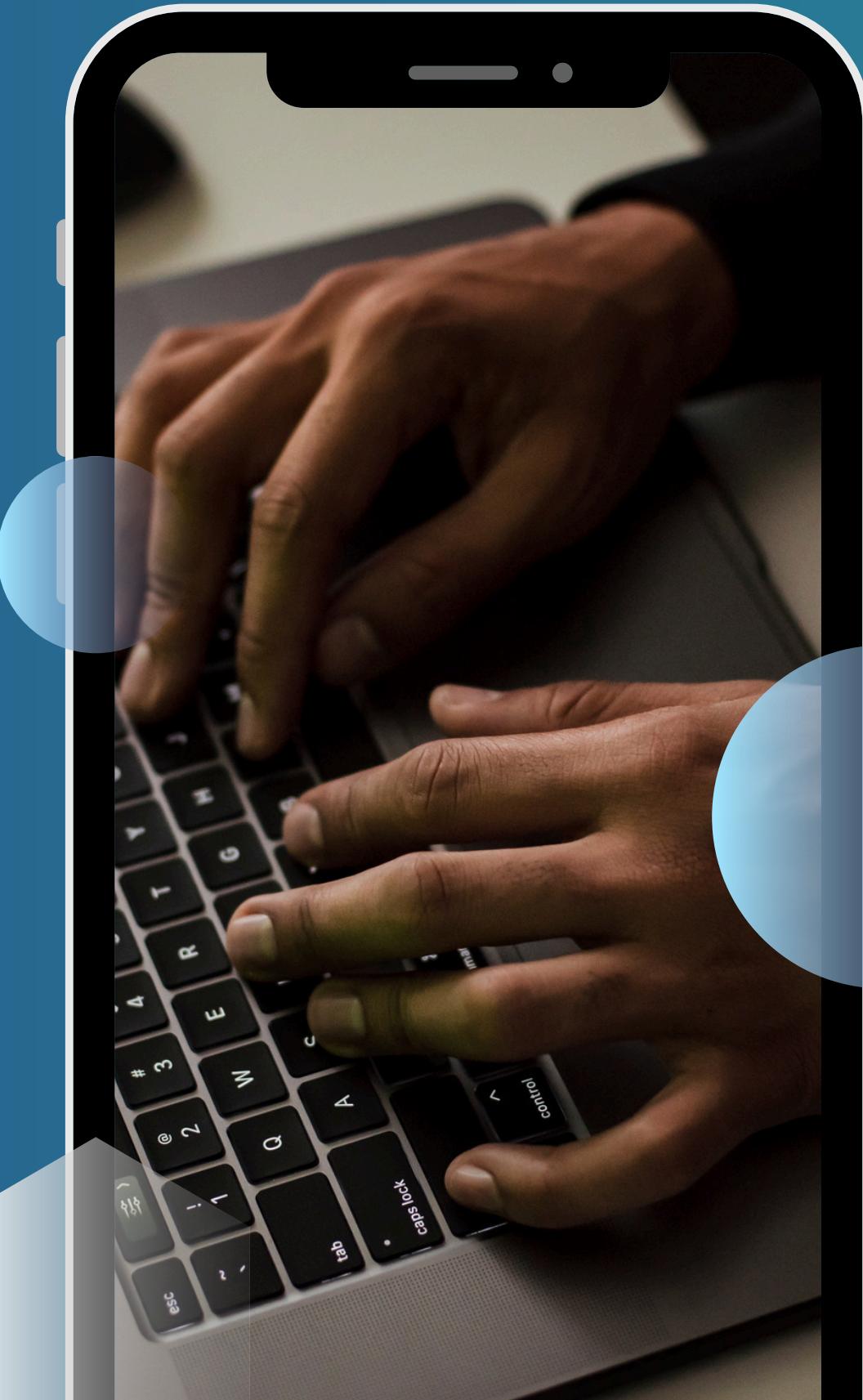
Source Visual B I Header 3 Format Insert Table

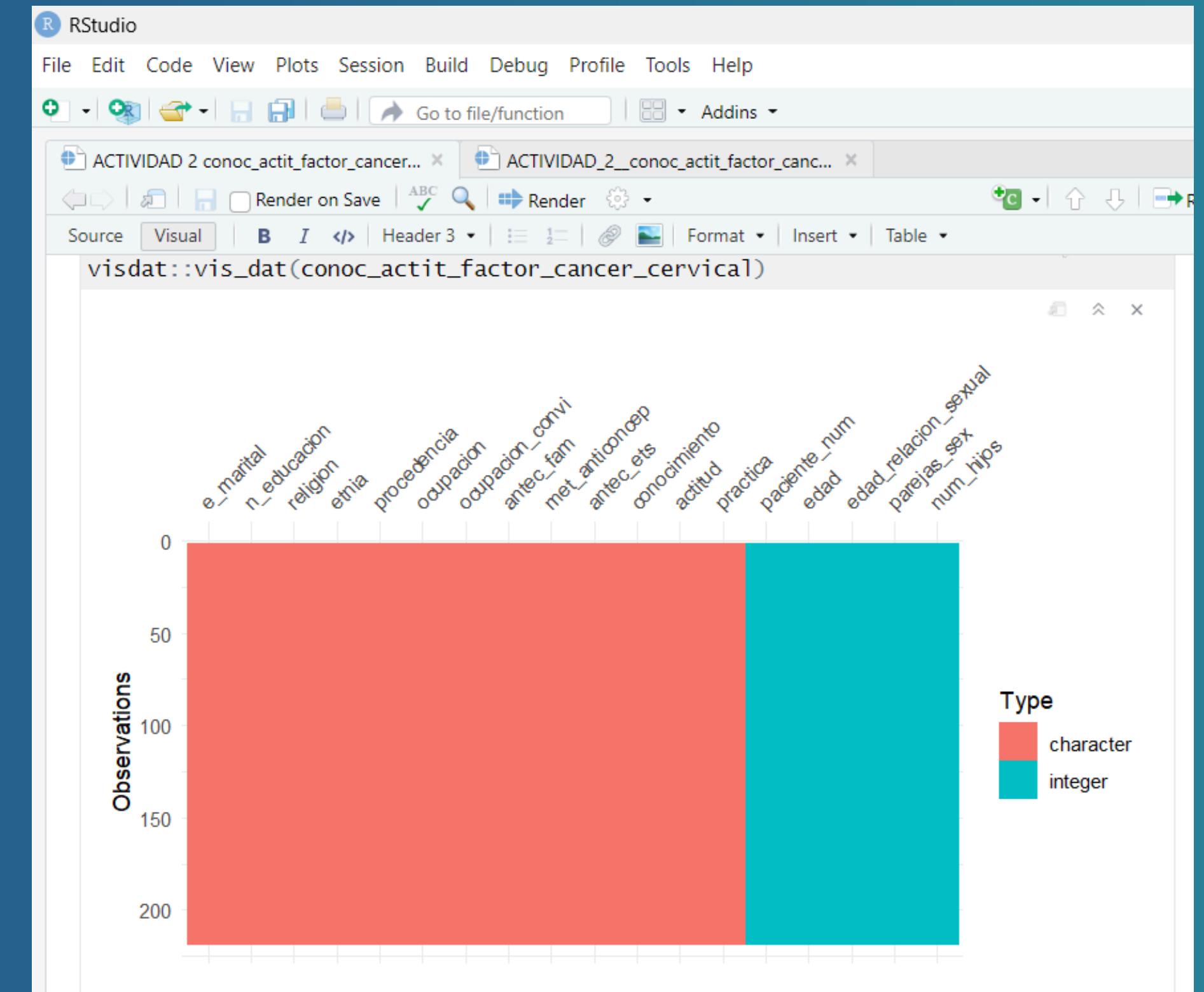
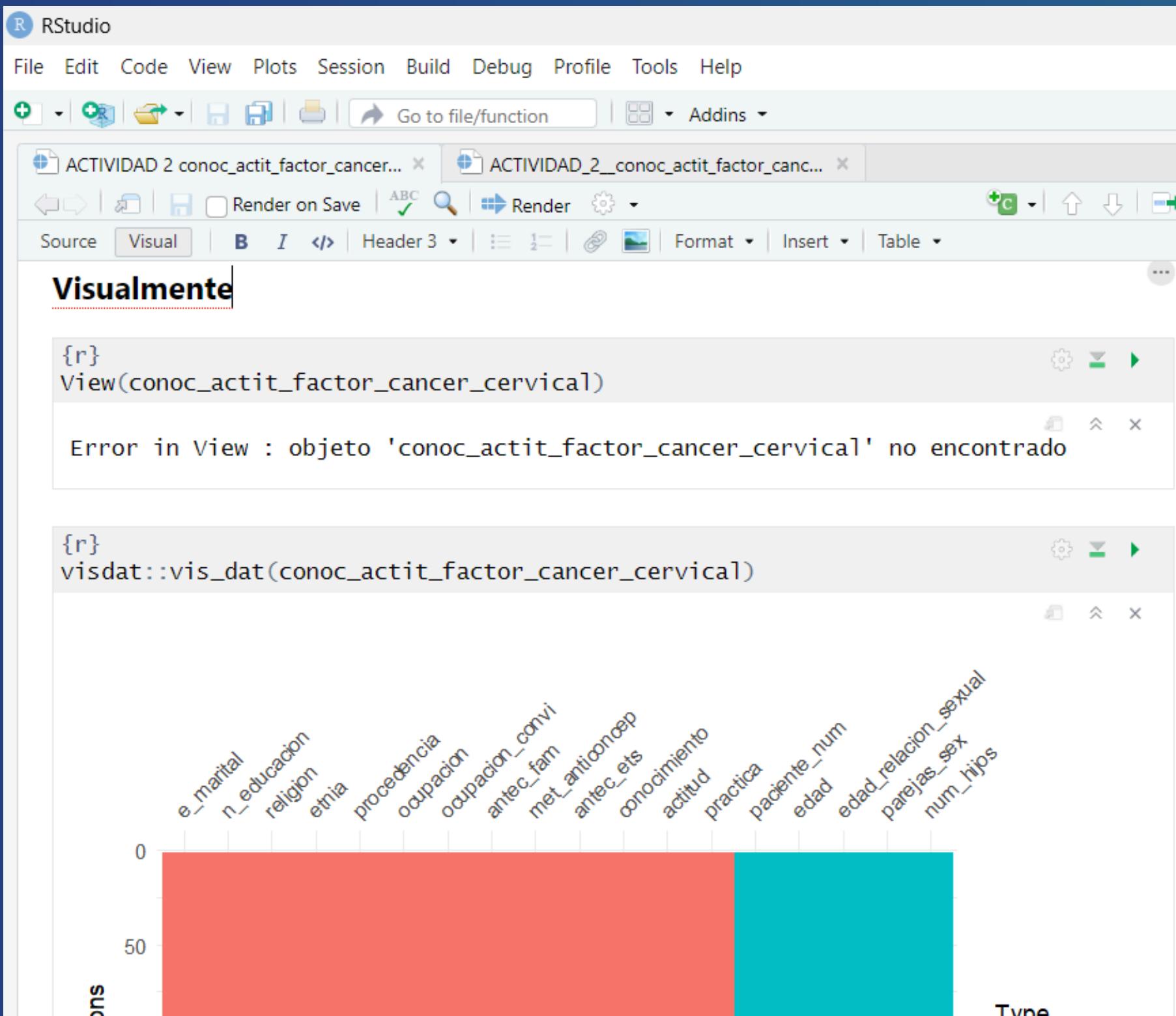
Resumen por variable

```
{r}
summary(conoc_actit_factor_cancer_cervical)
```

	paciente_num	edad	e_marital	n_educacion
religion				
Min.	: 1.00	Min. :25.00	Length:218	Length:218
Length	:218			
1st Qu.	: 55.25	1st Qu.:38.25	Class :character	Class :character Class
Median	:109.50	Median :48.00	Mode :character	Mode :character Mode
Mean	:109.50	Mean :45.41		
3rd Qu.	:163.75	3rd Qu.:54.00		
Max.	:218.00	Max. :67.00		
etnia		procedencia	ocupacion	ocupacion_convi
Length	:218	Length:218	Length:218	Length:218
Class	:character	Class :character	Class :character	Class :character
Mode	:character	Mode :character	Mode :character	Mode :character
antec_fam		edad_relacion_sexual	parejas_sex	num_hijos
met_anticoncep		Min. :13.00	Min. : 0.000	Min. :0.00
Length	:218			

Visualmente





Limpieza de datos

Paso uno: corregir los nombres de variables.

Clean names es una función del paquete janitor

```
{r}
conoc_actit_factor_cancer_cervical_1 = clean_names
(conoc_actit_factor_cancer_cervical)
```

Nota el contraste (la función `names()` imprime los nombres de columnas de un *dataset*)

```
{r}
names(conoc_actit_factor_cancer_cervical)
```

[1] "paciente_num"	"edad"	"e_marital"
[4] "n_educacion"	"religion"	"etnia"
[7] "procedencia"	"ocupacion"	"ocupacion_convi"
[10] "antec_fam"	"edad_relacion_sexual"	"parejas_sex"
[13] "num_hijos"	"met_anticoncep"	"antec_ets"
[16] "conocimiento"	"actitud"	"practica"

```
{r}  
names(conoc_actit_factor_cancer_cervical_1)
```

```
[1] "paciente_num"          "edad"           "e_marital"  
[4] "n_educacion"          "religion"        "etnia"  
[7] "procedencia"          "ocupacion"       "ocupacion_convi"  
[10] "antec_fam"            "edad_relacion_sexual" "parejas_sex"  
[13] "num_hijos"             "met_anticoncep"  "antec_ets"  
[16] "conocimiento"         "actitud"         "practica"
```

Paso dos: convertir celdas vacías a NA

```
{r}  
conoc_actit_factor_cancer_cervical_2 = mutate_if  
(conoc_actit_factor_cancer_cervical_1, is.character, list(~na_if(.,"")))
```

Paso tres: eliminar columnas o filas vacías.

```
{r}  
conoc_actit_factor_cancer_cervical_3 = remove_empty  
(conoc_actit_factor_cancer_cervical_2, which = c("rows", "cols"))
```

Optimizando el código

Corregir nombres, celdas vacías a NA y eliminar columnas o filas vacías.

```
{r}  
conoc_actit_factor_cancer_cervical_1 = conoc_actit_factor_cancer_cervical |>  
  clean_names() |>  
  mutate_if(is.character, list(~ na_if(.,"")))|>  
  remove_empty(which = c("rows", "cols"))
```



Paso 4: corregir errores ortográficos o valores inválidos

Inspección tabular

{r}

```
conoc_actit_factor_cancer_cervical_1 |> count(procedencia) # Cambia de variable  
categórica
```

Description: df [2 x 2]

procedencia	n
rural	107
urbano	111

<int>

2 rows

{r}

```
conoc_actit_factor_cancer_cervical_1 |> count(n_educacion) # Cambia de variable  
categórica
```

Description: df [3 x 2]

n_educacion	n
primaria	5
secundaria	53
superior	160

<int>

3 rows

```
{r}
conoc_actit_factor_cancer_cervical_1 |> count(etnia) # Cambia de variable
categórica
```

Description: df [3 x 2]

etnia	n
blanco	40
mestizo	163
otro	15

3 rows

Transformando de data.frame a as tibble

```
{r}
conoc_actit_factor_cancer_cervical_2 = as_tibble(
  conoc_actit_factor_cancer_cervical_1)
```

Corregir errores ortográficos usando `mutate()` y `case_when()`

```
{r}
conoc_actit_factor_cancer_cervical_3 = conoc_actit_factor_cancer_cervical_2 |>
  mutate(e_marital = case_when(
    e_marital == "casada" ~ "Casada",
    e_marital == "soltera" ~ "Soltera",
    e_marital == "conviviente" ~ "Conviviente",
    e_marital == "viuda" ~ "Viuda",
    TRUE ~ e_marital))
```

```
{r}
conoc_actit_factor_cancer_cervical_3 = conoc_actit_factor_cancer_cervical_2 |>
  mutate(religion = case_when(
    religion == "catolico" ~ "Catolico",
    religion == "catolico" ~ "católico",
    religion == "ninguna" ~ "Ninguna",
    religion == "evangelista" ~ "Evangelista",
    TRUE ~ religion))
```



Paso 5: Colapsar una variable categórica en menos niveles

Un vistazo a la variable de interés

```
{r}
conoc_actit_factor_cancer_cervical_3 |> count(conocimiento)
```

A tibble: 4 × 2

conocimiento	n
<chr>	<int>
alto	85
bajo	60
medio	67
no conoce	6

4 rows

Colapsar a dos categorías

```
{r}
library(dplyr)

conoc_actit_factor_cancer_cervical_4 = conoc_actit_factor_cancer_cervical_3 |>
  mutate(conocimiento_2 = case_when(
    conocimiento %in% c("no conoce") ~ "no tiene",
    conocimiento %in% c("bajo", "medio", "alto") ~ "si tiene",
    TRUE ~ conocimiento)
  )
```

Comprobando el cambio

{r}

conoc_actit_factor_cancer_cervical_4 |> count(conocimiento_2)

A tibble: 2 × 2

conocimiento_2	n
<chr>	<int>
no tiene	6
si tiene	212

2 rows

Paso 6: Transformar una variable

Transformación a logaritmo

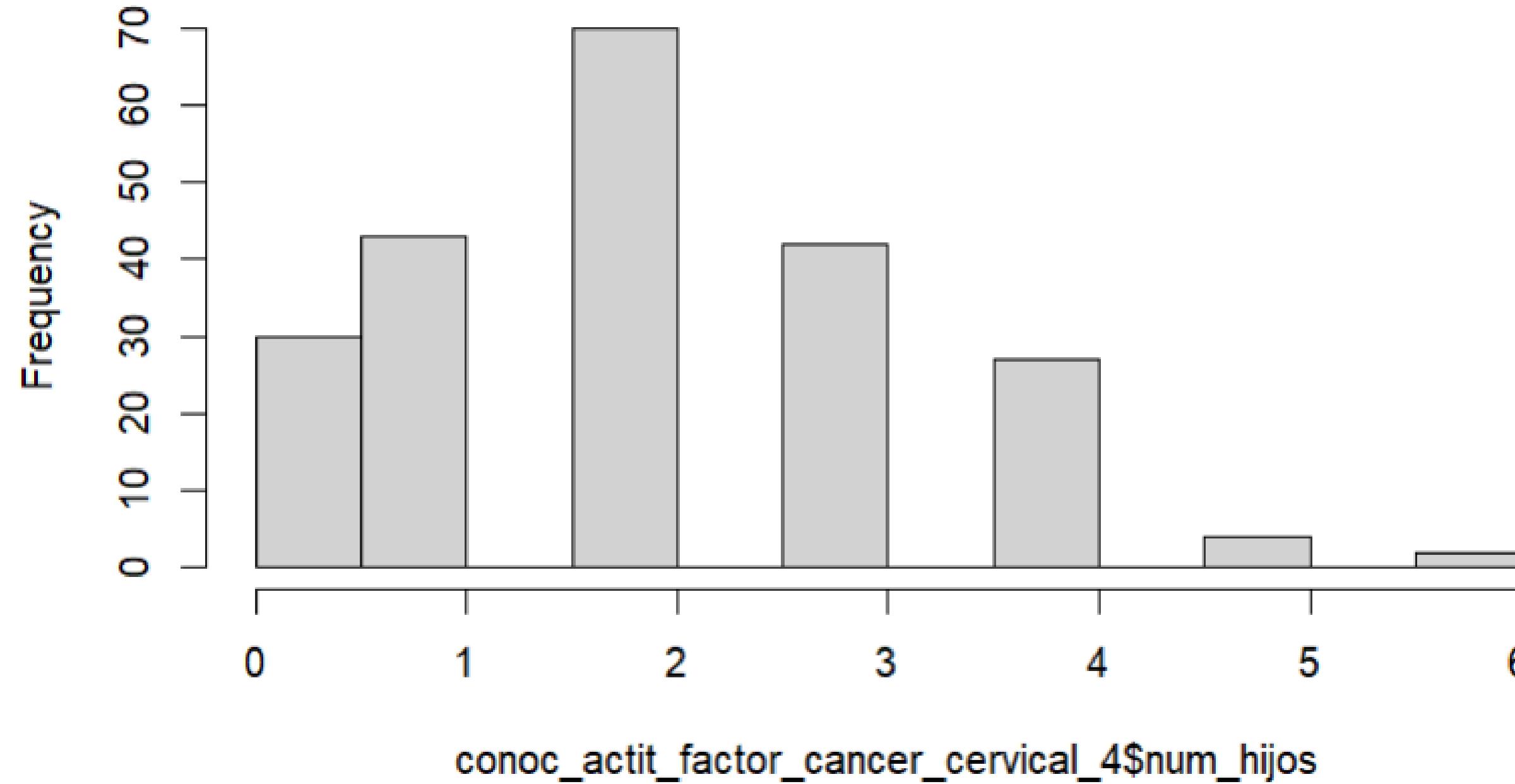
```
{r}  
summary(conoc_actit_factor_cancer_cervical_4$num_hijos)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	1.00	2.00	2.06	3.00	6.00

```
{r}
```

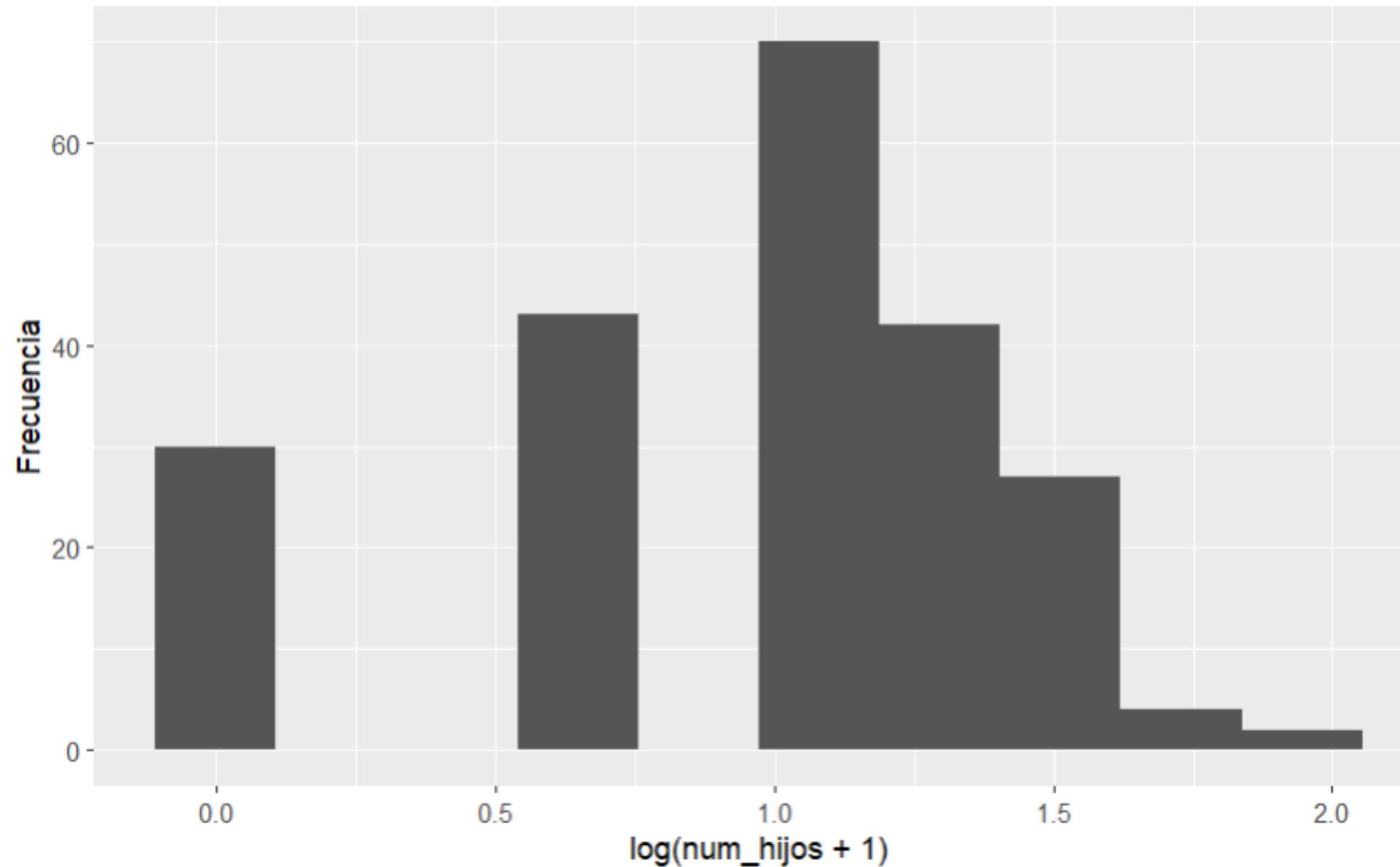
```
hist(conoc_actit_factor_cancer_cervical_4$num_hijos)
```

Histogram of conoc_actit_factor_cancer_cervical_4\$num_hijos



```
{r}
conoc_actit_factor_cancer_cervical_5 <- conoc_actit_factor_cancer_cervical_4 |>
  mutate(num_hijos = (log(num_hijos + 1)))
```

```
{r}
conoc_actit_factor_cancer_cervical_5 |>
  ggplot(aes(x = num_hijos)) +
  geom_histogram(bins = 10) +
  labs(y = "Frecuencia", x = "log(num_hijos + 1)")
```



Transformación a binario

```
{r}
conoc_actit_factor_cancer_cervical_6 = conoc_actit_factor_cancer_cervical_5 |>
  mutate(
    parejas_sex_c = case_when(
      parejas_sex < 10 ~ "< 10",
      parejas_sex >= 10 ~ ">= 10"
    )
  ) |>
  mutate(parejas_sex_c = factor(parejas_sex, levels = "< 10", ">= 10"))
```

Transformando valores a valores perdidos usando la función `na_if()`

```
{r}
conoc_actit_factor_cancer_cervical_7 = conoc_actit_factor_cancer_cervical_6 |>
  mutate(parejas_sex = na_if(parejas_sex, -7))
```

Transformando valores a valores perdidos usando la función `case_when()`

```
{r}
conoc_actit_factor_cancer_cervical_7 = conoc_actit_factor_cancer_cervical_6 |>
  mutate(edad = case_when(edad %in% c(3, 999) ~ NA,
                         TRUE ~ edad))
```

Paso 7: Renombrar una variable

Imprimir los nombres. ¿Cuáles necesitan cambio?

```
{r}  
names(conoc_actit_factor_cancer_cervical_7)
```

```
[1] "paciente_num"           "edad"          "e_marital"  
[4] "n_educacion"            "religion"       "etnia"  
[7] "procedencia"            "ocupacion"     "ocupacion_convi"  
[10] "antec_fam"              "edad_relacion_sexual" "parejas_sex"  
[13] "num_hijos"               "met_anticoncep" "antec_ets"  
[16] "conocimiento"            "actitud"        "practica"  
[19] "conocimiento_2"          "parejas_sex_c"
```

Cambiando un nombre de variables

```
{r}  
conoc_actit_factor_cancer_cervical_8 <- conoc_actit_factor_cancer_cervical_7 |>  
  rename(Número_de_paciente = paciente_num)
```

Varios a la vez

```
{r}
conoc_actit_factor_cancer_cervical_8 <- conoc_actit_factor_cancer_cervical_7 |>
  rename(Número_de_paciente = paciente_num,
         Edad = edad,
         Estado_marital = e_marital,
         Nivel_de_educación = n_educacion,
         Procedencia = procedencia ,
         Antecedentes_fam = antec_fam,
         Número_de_paciente = paciente_num,
         Conocimiento = conocimiento,
         Conocimiento_2 = conocimiento_2,
         Religión = religion,
         Ocupación = ocupacion,
         Edad_relacion_sexual = edad_relacion_sexual,
         Método_anticoceptivo = met_anticoncep,
         Actitud = actitud,
         Parejas_sex_c = parejas_sex_c,
         Etnia = etnia,
         Ocupación_convi = ocupacion_convi ,
         Parejas_sex = parejas_sex,
         Antecedentes_ets = antec_ets,
         Practica = practica)
```

Comprobando

```
{r}  
names(conoc_actit_factor_cancer_cervical_8)
```

```
[1] "Número_de_paciente"      "Edad"                  "Estado_marital"  
[4] "Nivel_de_educación"       "religion"               "etnia"  
[7] "procedencia"              "ocupacion"              "ocupacion_convi"  
[10] "antec_fam"                "edad_relacion_sexual" "parejas_sex"  
[13] "num_hijos"                "met_anticoncep"        "antec_ets"  
[16] "conocimiento"             "actitud"                "practica"  
[19] "conocimiento_2"            "parejas_sex_c"
```