lab1

Cristopher Barrios, Josue Sagastume

21/7/2023

1. Haga una exploración rápida de sus datos.

```
df = read.csv("./risk_factors_cervical_cancer.csv")
```

```
# Crear un nuevo DataFrame seleccionando solo las columnas deseadas
nuevo_df <- data.frame(
  Age = df$Age,
  Number_of_sexual_partners = as.numeric(df$Number.of.sexual.partners),
  First_sexual_intercourse = as.numeric(df$First.sexual.intercourse),
  Num_of_pregnancies = as.numeric(df$Num.of.pregnancies),
  Smokes = as.numeric(df$Smokes),
  Hormonal_Contraceptives = as.numeric(df$Hormonal.Contraceptives),
  IUD = as.numeric(df$IUD),
  STDs = as.numeric(df$STDs),
  Dx_Cancer = df$Dx.Cancer,
  Dx_HPV = df$Dx.HPV,
  Biopsy = df$Biopsy
)

# Opcionalmente, puedes renombrar las columnas si deseas utilizar nombres más cortos
colnames(nuevo_df) <- c("Age", "Number_of_sexual_partners", "First_sexual_intercourse",
                        "Num_of_pregnancies", "Smokes", "Hormonal_Contraceptives",
                        "IUD", "STDs", "Dx_Cancer", "Dx_HPV", "Biopsy")
```

```
library(readr)
library(dplyr)
```

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

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

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

```
data <- read_csv('risk_factors_cervical_cancer.csv')
```

```
## Rows: 858 Columns: 36
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (26): Number of sexual partners, First sexual intercourse, Num of pregna...
## dbl (10): Age, STDs: Number of diagnosis, Dx:Cancer, Dx:CIN, Dx:HPV, Dx, Hin...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
data <- data %>% mutate_all(~replace(., . == "?", NA))


missing_values <- colSums(is.na(data))


numerical_stats <- data %>% summarise(across(where(is.numeric), list(mean = mean, std_dev = sd, min = mi


categorical_stats <- data %>% summarise(across(where(is.character), list(unique = ~length(unique(.)), mo
```

```
missing_values
```

```
##                                 Age          Number of sexual partners
##                                   0                                 26
##           First sexual intercourse                Num of pregnancies
##                                   7                                 56
##                             Smokes                    Smokes (years)
##                                  13                                 13
##                Smokes (packs/year)            Hormonal Contraceptives
##                                  13                                108
##    Hormonal Contraceptives (years)                                IUD
##                                 108                                117
##                         IUD (years)                               STDs
##                                 117                                105
##                       STDs (number)               STDs:condylomatosis
##                                 105                                105
##       STDs:cervical condylomatosis       STDs:vaginal condylomatosis
##                                 105                                105
## STDs:vulvo-perineal condylomatosis                     STDs:syphilis
##                                 105                                105
##   STDs:pelvic inflammatory disease               STDs:genital herpes
##                                 105                                105
##         STDs:molluscum contagiosum                          STDs:AIDS
##                                 105                                105
##                           STDs:HIV                  STDs:Hepatitis B
##                                 105                                105
##                           STDs:HPV          STDs: Number of diagnosis
##                                 105                                  0
##   STDs: Time since first diagnosis    STDs: Time since last diagnosis
```

```
##                                 787                                 787
##                          Dx:Cancer                              Dx:CIN
##                                   0                                   0
##                             Dx:HPV                                  Dx
##                                   0                                   0
##                           Hinselmann                            Schiller
##                                   0                                   0
##                            Citology                              Biopsy
##                                   0                                   0
```

```
#numerical_stats
#categorical_stats
```

Este bloque de texto muestra un resumen de algunas variables del conjunto de datos, donde se presenta la cantidad de valores distintos y la cantidad de registros en los que aparecen esos valores.

Number of sexual partners: La variable "Number of sexual partners" representa la cantidad de parejas sexuales. Se observa que hay 26 registros con un número específico de parejas sexuales.

STDs: Esta variable indica si la persona tiene alguna infección de transmisión sexual (ITS). Hay 105 registros con valores específicos sobre la presencia o ausencia de ITS.

```
summary(nuevo_df)
```

```
##       Age        Number_of_sexual_partners First_sexual_intercourse
##  Min.   :13.00   Min.   : 1.000            Min.   :10
##  1st Qu.:20.00   1st Qu.: 2.000            1st Qu.:15
##  Median :25.00   Median : 2.000            Median :17
##  Mean   :26.82   Mean   : 2.528            Mean   :17
##  3rd Qu.:32.00   3rd Qu.: 3.000            3rd Qu.:18
##  Max.   :84.00   Max.   :28.000            Max.   :32
##                  NA's   :26                NA's   :7
##  Num_of_pregnancies     Smokes       Hormonal_Contraceptives      IUD
##  Min.   : 0.000     Min.   :0.0000   Min.   :0.0000          Min.   :0.000
##  1st Qu.: 1.000     1st Qu.:0.0000   1st Qu.:0.0000          1st Qu.:0.000
##  Median : 2.000     Median :0.0000   Median :1.0000          Median :0.000
##  Mean   : 2.276     Mean   :0.1456   Mean   :0.6413          Mean   :0.112
##  3rd Qu.: 3.000     3rd Qu.:0.0000   3rd Qu.:1.0000          3rd Qu.:0.000
##  Max.   :11.000     Max.   :1.0000   Max.   :1.0000          Max.   :1.000
##  NA's   :56         NA's   :13       NA's   :108             NA's   :117
##       STDs          Dx_Cancer           Dx_HPV            Biopsy
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.0000   Median :0.00000   Median :0.00000   Median :0.0000
##  Mean   :0.1049   Mean   :0.02098   Mean   :0.02098   Mean   :0.0641
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000
##  NA's   :105
```

```
# Ver las primeras filas del nuevo DataFrame
head(nuevo_df)
```

```
##   Age Number_of_sexual_partners First_sexual_intercourse Num_of_pregnancies
```

```
## 1  18                          4                          15                  1
## 2  15                          1                          14                  1
## 3  34                          1                          NA                  1
## 4  52                          5                          16                  4
## 5  46                          3                          21                  4
## 6  42                          3                          23                  2
##   Smokes Hormonal_Contraceptives IUD STDs Dx_Cancer Dx_HPV Biopsy
## 1      0                        0   0    0         0      0      0
## 2      0                        0   0    0         0      0      0
## 3      0                        0   0    0         0      0      0
## 4      1                        1   0    0         1      1      0
## 5      0                        1   0    0         0      0      0
## 6      0                        0   0    0         0      0      0
```

2. Diga el tipo de cada una de las variables del dataset (cualitativa o categórica, cuantitativa continua, cuantitativa discreta

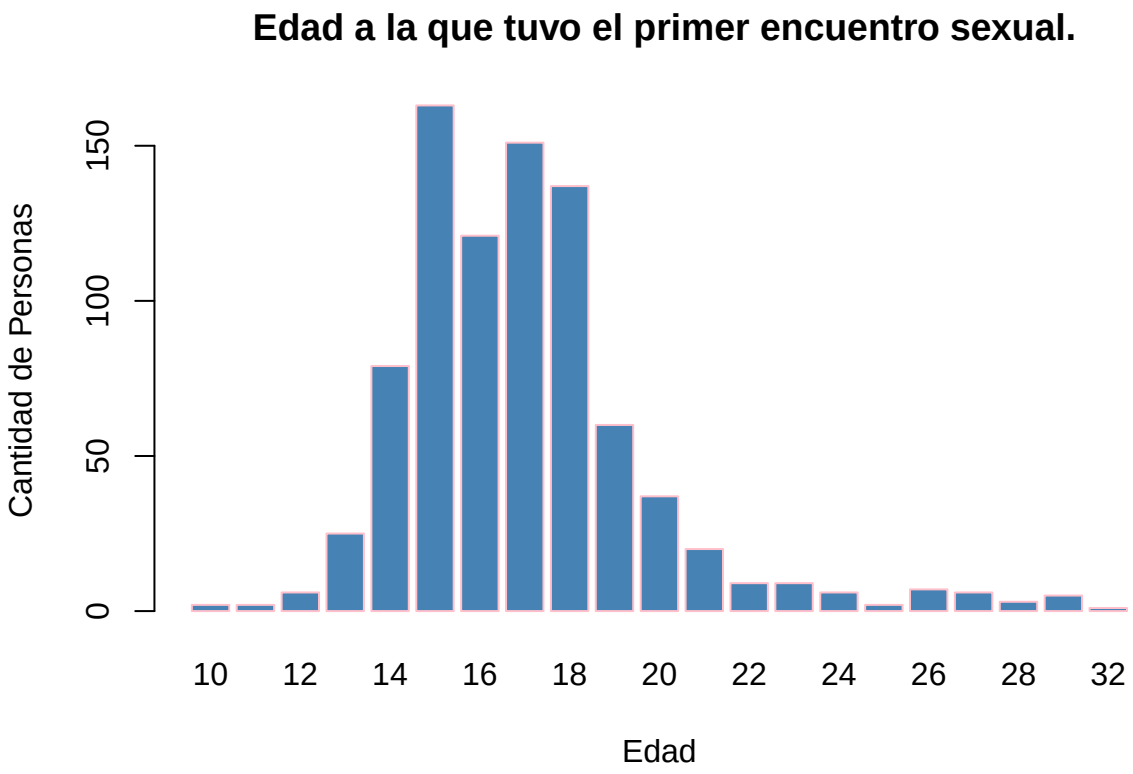
Cuantitativas discretas: STDs (number), Age, Number of sexual partners, STDs: Number of diagnosis, Num of pregnancies

Cuantitativas continuas: First sexual intercourse, STDs: Time since last diagnosis, STDs: Time since first diagnosis, Smokes (years), Smokes (packs/year), IUD (years), Hormonal Contraceptives (years)

Cualitativas: STDs:cervical condylomatosis, Schiller, STDs:HPV, IUD, STDs:vulvo-perineal condylomatosis, STDs, STDs:syphilis, Dx:HPV, Dx:Cancer, STDs:HIV, STDs:molluscum contagiosum, STDs:condylomatosis, Dx:CIN, Hinselmann, STDs:Hepatitis B, Hormonal Contraceptives, Biopsy, Citology, STDs:genital herpes, STDs:vaginal condylomatosis, Dx, STDs:AIDS, Smokes

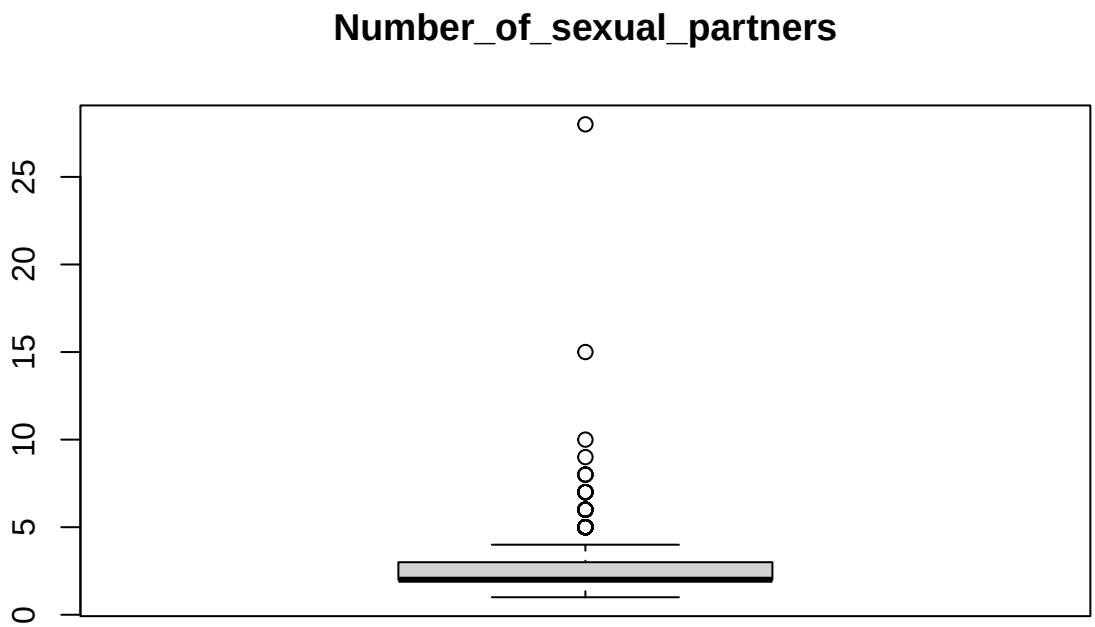
3. Incluya los gráficos exploratorios siendo consecuentes con el tipo de variable que están representando.

```
barplot(table(nuevo_df$First_sexual_intercourse), main = " Edad a la que tuvo el primer encuentro sexual
```



La mayoria de las personas en este conjunto de datos tienensu primer encuentro sexual entre los 14 años hasta los 20

```
boxplot(nuevo_df$Number_of_sexual_partners, main = "Number_of_sexual_partners", xlab = "Registro por añ
```

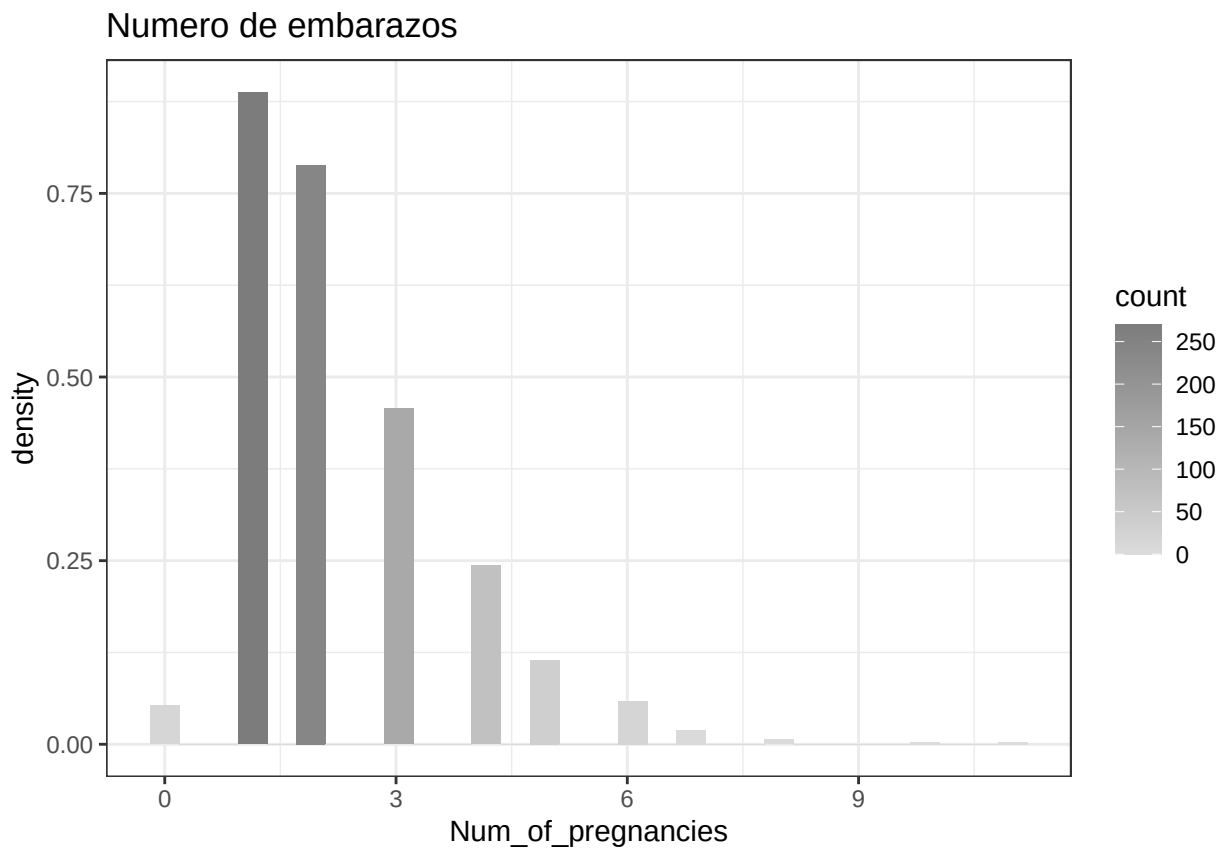


Registro por año

Pues se puede observar muy pocos puntos atipicos pero existe una persona que ha tenido aproximadamente 30 encuentros sexuales, la mayoria tienen mas o menos 4 encuentros sexuales, y no pasan de 15.

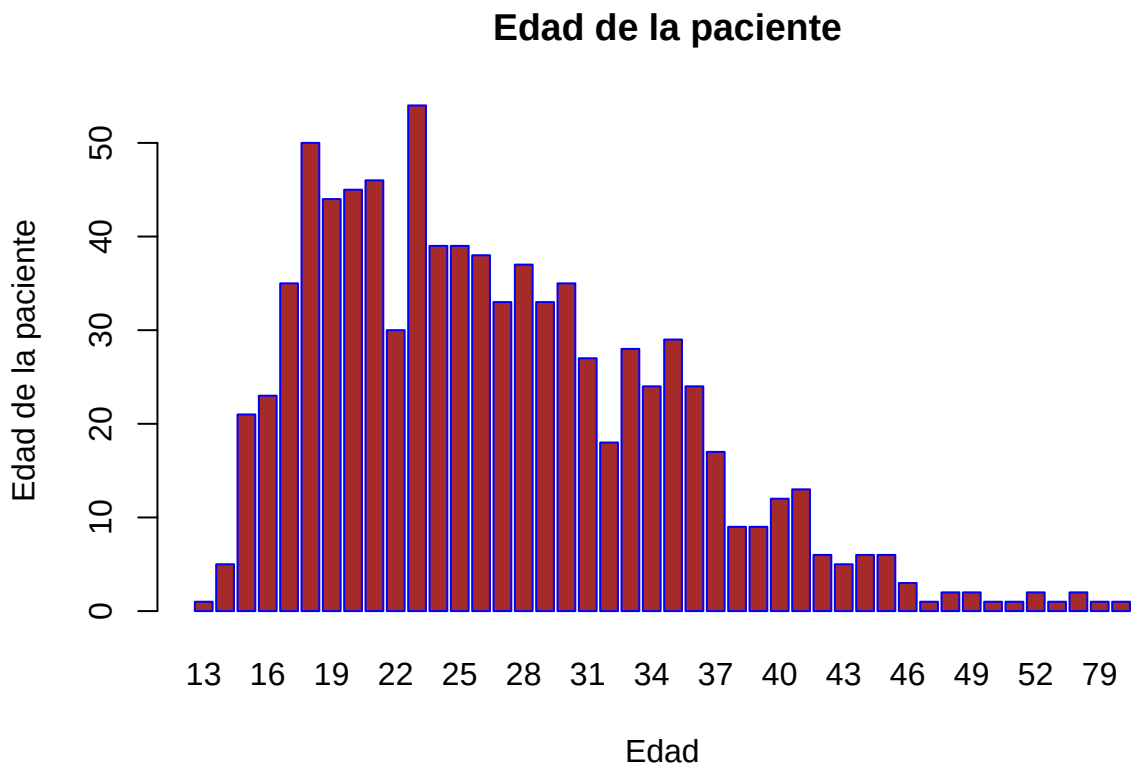
```
# Cargar el paquete ggplot2
library(ggplot2)

ggplot(data = nuevo_df, aes(x = Num_of_pregnancies)) +
  geom_histogram(aes(y=..density.., fill = ..count..))+
  scale_fill_gradient(low = "#DCDCDC", high = "#7C7C7C")+
  stat_function(fun = dnorm, colour = "firebrick",
                args = list(mean = mean(nuevo_df$Num_of_pregnancies),
                            sd = sd(nuevo_df$Num_of_pregnancies)))+
  ggtitle("Numero de embarazos ") +
  theme_bw()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



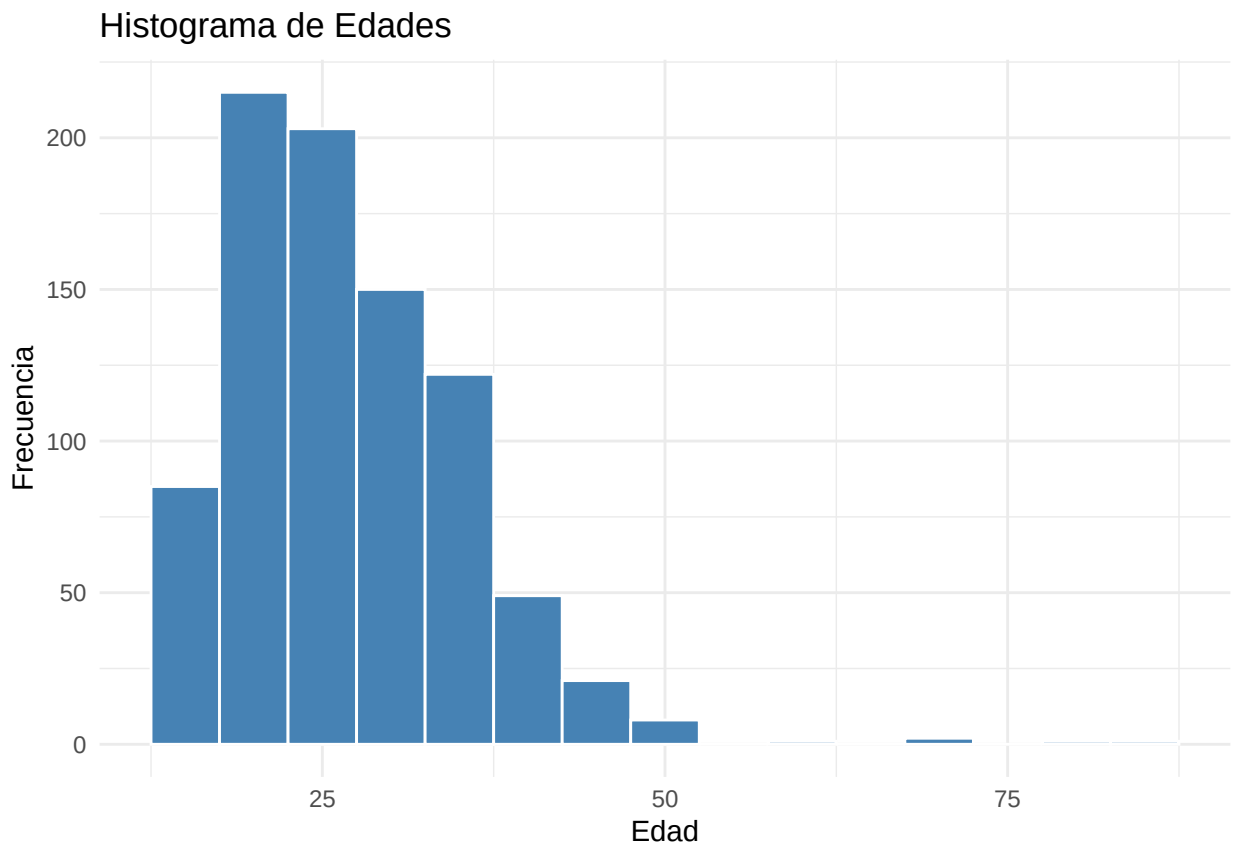
En el eje horizontal, se encuentran los diferentes valores enteros de embarazos, mientras que en el eje vertical se muestra la frecuencia con la que cada valor ocurre en la muestra. Cada barra del histograma representa un rango de valores de embarazos y su altura indica la cantidad de veces que se observa dicho rango en el conjunto de datos. A través de este gráfico, podemos identificar patrones en la frecuencia de embarazos y tener una visión general de la tendencia de esta variable en la población estudiada

```
barplot(table(nuevo_df$Age), main = " Edad de la paciente", xlab = "Edad", ylab = "Edad de la paciente"
```



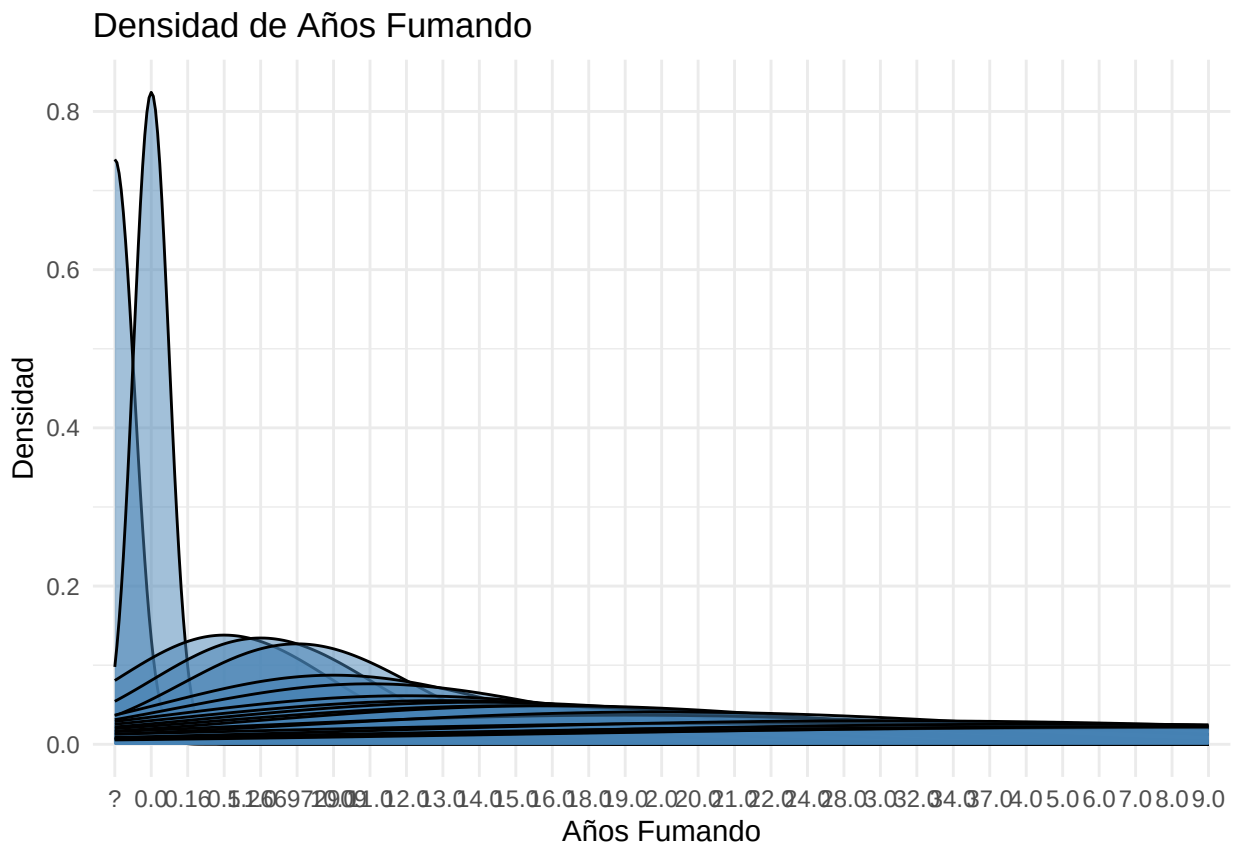
Las edades son muy variadas, parece que la edad en la que mas hay son de 22 añospero la gran mayoria casi en el mismo nivel entre 16 años hasta los 37 años.

```
ggplot(df, aes(x = Age)) +
  geom_histogram(binwidth = 5, fill = "steelblue", color = "white") +
  labs(title = "Histograma de Edades", x = "Edad", y = "Frecuencia") +
  theme_minimal()
```



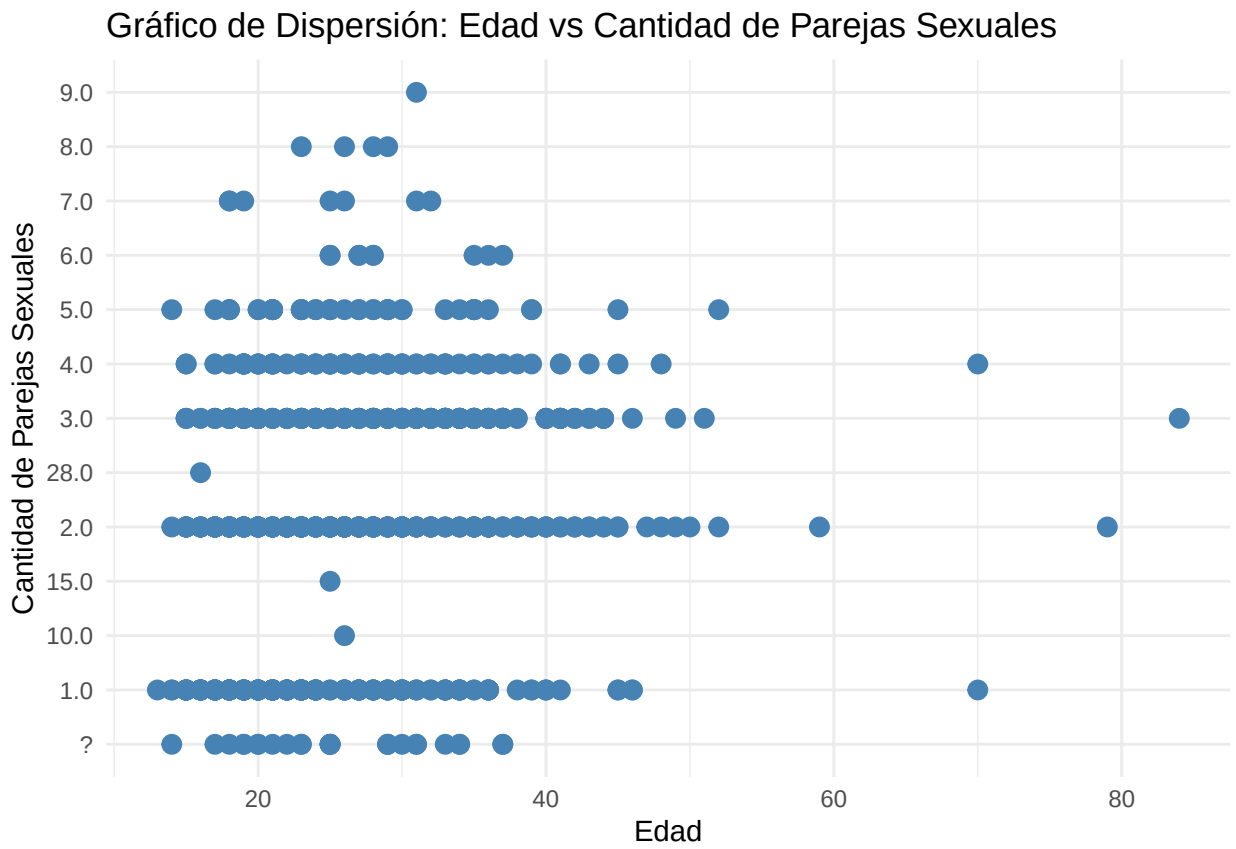
Este histograma de edades podemos observar que llega gente muy joven a este sitio por diferentes enfermedades, y la adulta casi no llega.

```
ggplot(df, aes(x = `Smokes..years.`)) +
  geom_density(fill = "steelblue", alpha = 0.5) +
  labs(title = "Densidad de Años Fumando", x = "Años Fumando", y = "Densidad") +
  theme_minimal()
```



Se puede observar que mientras mas pasan los años las personas dejan de fumar despues de cierto tiempo, esto puede ser debido al diagnostico que han recibido.

```
ggplot(df, aes(x = `Age`, y = `Number.of.sexual.partners`)) +
  geom_point(color = "steelblue", size = 3) +
  labs(title = "Gráfico de Dispersión: Edad vs Cantidad de Parejas Sexuales",
       x = "Edad", y = "Cantidad de Parejas Sexuales") +
  theme_minimal()
```



Se puede observar que la mayoria son personas de menos de 40 años y que aproximadamente a los 30 años se tienen mas encuentros sexuales.

4. Aísle las variables numéricas de las categóricas, haga un análisis de correlación entre las mismas.

```
df <- read.csv("./risk_factors_cervical_cancer.csv")

library(dplyr)
library(corrr)



library(FactoMineR)
library(fpc)
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```r
#df <- read.csv("./data/Ejemplo.csv", stringsAsFactors = FALSE)

#¿Reemplazar '?' con NA (valores faltantes)
df[df == "?"] <- NA

# ¿onvertir las columnas especificadas a numéricas
numeric_columns <- c('Age','Number.of.sexual.partners', 'First.sexual.intercourse', 'Num.of.pregnancies
                     'Smokes..years.', 'Smokes..packs.year.', 'Hormonal.Contraceptives..years.',
                     'IUD..years.', 'STDs..number.')

df[numeric_columns] <- lapply(df[numeric_columns], as.numeric)

# ¿ Seleccionar solo las columnas numéricas y crear la matriz de correlación
numeric_data <- df[, sapply(df, is.numeric)]
#correlation_matrix <- correlate(numeric_data, method = "pearson")

# Imprimir la matriz de correlación
#print(correlation_matrix)

cor(numeric_data[,-1],use = "pairwise.complete.obs")
```

```
##                                   Number.of.sexual.partners
## Number.of.sexual.partners                       1.000000000
## First.sexual.intercourse                       -0.150168584
## Num.of.pregnancies                              0.079080651
## Smokes..years.                                  0.186932306
## Smokes..packs.year.                             0.182066838
## Hormonal.Contraceptives..years.                 0.019569271
## IUD..years.                                     0.004453860
## STDs..number.                                   0.041441676
## STDs..Number.of.diagnosis                       0.051899624
## Dx.Cancer                                       0.022316365
## Dx.CIN                                          0.015693534
## Dx.HPV                                          0.027272910
## Dx                                              0.022991951
## Hinselmann                                     -0.039847408
## Schiller                                       -0.008966946
## Citology                                        0.021857561
## Biopsy                                         -0.001442452
##                                   First.sexual.intercourse Num.of.pregnancies
## Number.of.sexual.partners                     -0.150168584        0.079080651
## First.sexual.intercourse                       1.000000000       -0.060732749
## Num.of.pregnancies                            -0.060732749        1.000000000
## Smokes..years.                                -0.058834446        0.180330830
## Smokes..packs.year.                           -0.056755203        0.100903857
## Hormonal.Contraceptives..years.                0.008307764        0.224790029
## IUD..years.                                   -0.026503290        0.154987170
## STDs..number.                                  0.006745144        0.001743902
## STDs..Number.of.diagnosis                     -0.013331999        0.034153312
## Dx.Cancer                                      0.067288505        0.035149052
## Dx.CIN                                        -0.032627759        0.010985120
## Dx.HPV                                         0.043969587        0.046788003
## Dx                                             0.035754625        0.021335813
```

```
## Hinselmann                                  -0.016548708        0.040438701
## Schiller                                     0.003494755        0.092017287
## Citology                                    -0.010973152       -0.030036158
## Biopsy                                       0.007264369        0.046415929
##                                 Smokes..years. Smokes..packs.year.
## Number.of.sexual.partners          0.186932306         0.182066838
## First.sexual.intercourse          -0.058834446        -0.056755203
## Num.of.pregnancies                 0.180330830         0.100903857
## Smokes..years.                     1.000000000         0.724115662
## Smokes..packs.year.                0.724115662         1.000000000
## Hormonal.Contraceptives..years.    0.050979204         0.041277637
## IUD..years.                        0.040219705         0.016583893
## STDs..number.                      0.098771570         0.032657875
## STDs..Number.of.diagnosis          0.084646697         0.033356590
## Dx.Cancer                          0.056234118         0.111572023
## Dx.CIN                            -0.030968273        -0.021128434
## Dx.HPV                             0.058847278         0.113515869
## Dx                                -0.049926012        -0.034062554
## Hinselmann                         0.072251281         0.027043544
## Schiller                           0.095890414         0.018191236
## Citology                          -0.006827071         0.004665534
## Biopsy                             0.062044187         0.024881930
##                                 Hormonal.Contraceptives..years.   IUD..years.
## Number.of.sexual.partners                         0.0195692714  0.0044538599
## First.sexual.intercourse                          0.0083077641 -0.0265032899
## Num.of.pregnancies                                0.2247900288  0.1549871702
## Smokes..years.                                    0.0509792041  0.0402197053
## Smokes..packs.year.                               0.0412776368  0.0165838932
## Hormonal.Contraceptives..years.                   1.0000000000  0.0004828029
## IUD..years.                                       0.0004828029  1.0000000000
## STDs..number.                                    -0.0070548347  0.0156663746
## STDs..Number.of.diagnosis                        -0.0384586779  0.0079025189
## Dx.Cancer                                         0.0547117808  0.0981127316
## Dx.CIN                                            0.0032731678  0.0180006866
## Dx.HPV                                            0.0632285501  0.0336468693
## Dx                                               -0.0134461126  0.1119927697
## Hinselmann                                        0.0389448201  0.0079940157
## Schiller                                          0.0792473114  0.0794152988
## Citology                                          0.0762631866  0.0027147030
## Biopsy                                            0.0793876331  0.0332753059
##                                 STDs..number. STDs..Number.of.diagnosis
## Number.of.sexual.partners          0.041441676               0.051899624
## First.sexual.intercourse           0.006745144              -0.013331999
## Num.of.pregnancies                 0.001743902               0.034153312
## Smokes..years.                     0.098771570               0.084646697
## Smokes..packs.year.                0.032657875               0.033356590
## Hormonal.Contraceptives..years.   -0.007054835              -0.038458678
## IUD..years.                        0.015666375               0.007902519
## STDs..number.                      1.000000000               0.897233276
## STDs..Number.of.diagnosis          0.897233276               1.000000000
## Dx.Cancer                         -0.018255589              -0.015422882
## Dx.CIN                            -0.009525770               0.008069606
## Dx.HPV                            -0.018255589              -0.015422882
## Dx                                -0.028340592              -0.002288585
```

```
## Hinselmann                          0.065349132              0.076786976
## Schiller                            0.120725311              0.130872847
## Citology                            0.060009521              0.055114464
## Biopsy                              0.098347334              0.097448921
##                                    Dx.Cancer       Dx.CIN      Dx.HPV
## Number.of.sexual.partners         0.02231637  0.015693534  0.02727291
## First.sexual.intercourse          0.06728851 -0.032627759  0.04396959
## Num.of.pregnancies                0.03514905  0.010985120  0.04678800
## Smokes..years.                    0.05623412 -0.030968273  0.05884728
## Smokes..packs.year.               0.11157202 -0.021128434  0.11351587
## Hormonal.Contraceptives..years.   0.05471178  0.003273168  0.06322855
## IUD..years.                       0.09811273  0.018000687  0.03364687
## STDs..number.                    -0.01825559 -0.009525770 -0.01825559
## STDs..Number.of.diagnosis        -0.01542288  0.008069606 -0.01542288
## Dx.Cancer                         1.00000000 -0.015071762  0.88650794
## Dx.CIN                           -0.01507176  1.000000000 -0.01507176
## Dx.HPV                            0.88650794 -0.015071762  1.00000000
## Dx                                0.66564706  0.606938678  0.61632710
## Hinselmann                        0.13426360 -0.021232519  0.13426360
## Schiller                          0.15781160  0.009119105  0.15781160
## Citology                          0.11344608 -0.023937652  0.11344608
## Biopsy                            0.16090497  0.113172334  0.16090497
##                                           Dx   Hinselmann     Schiller
## Number.of.sexual.partners         0.022991951 -0.039847408 -0.008966946
## First.sexual.intercourse          0.035754625 -0.016548708  0.003494755
## Num.of.pregnancies                0.021335813  0.040438701  0.092017287
## Smokes..years.                   -0.049926012  0.072251281  0.095890414
## Smokes..packs.year.              -0.034062554  0.027043544  0.018191236
## Hormonal.Contraceptives..years. -0.013446113  0.038944820  0.079247311
## IUD..years.                       0.111992770  0.007994016  0.079415299
## STDs..number.                    -0.028340592  0.065349132  0.120725311
## STDs..Number.of.diagnosis        -0.002288585  0.076786976  0.130872847
## Dx.Cancer                         0.665647059  0.134263602  0.157811599
## Dx.CIN                            0.606938678 -0.021232519  0.009119105
## Dx.HPV                            0.616327096  0.134263602  0.157811599
## Dx                                1.000000000  0.072214849  0.098952103
## Hinselmann                        0.072214849  1.000000000  0.650249194
## Schiller                          0.098952103  0.650249194  1.000000000
## Citology                          0.088739964  0.192467108  0.361486486
## Biopsy                            0.157606644  0.547416628  0.733203881
##                                      Citology       Biopsy
## Number.of.sexual.partners         0.021857561 -0.001442452
## First.sexual.intercourse         -0.010973152  0.007264369
## Num.of.pregnancies               -0.030036158  0.046415929
## Smokes..years.                   -0.006827071  0.062044187
## Smokes..packs.year.               0.004665534  0.024881930
## Hormonal.Contraceptives..years.   0.076263187  0.079387633
## IUD..years.                       0.002714703  0.033275306
## STDs..number.                     0.060009521  0.098347334
## STDs..Number.of.diagnosis         0.055114464  0.097448921
## Dx.Cancer                         0.113446079  0.160904975
## Dx.CIN                           -0.023937652  0.113172334
## Dx.HPV                            0.113446079  0.160904975
## Dx                                0.088739964  0.157606644
```

```
## Hinselmann                    0.192467108  0.547416628
## Schiller                      0.361486486  0.733203881
## Citology                      1.000000000  0.327466388
## Biopsy                        0.327466388  1.000000000
```

Esta tabla muestra los coeficientes de correlación, que varían entre -1 y 1, y reflejan la dirección y la intensidad de la asociación lineal entre pares de variables. Un coeficiente de correlación positivo cercano a 1 indica una relación directamente proporcional, lo que significa que a medida que una variable aumenta, la otra también tiende a aumentar. Por otro lado, un coeficiente negativo cercano a -1 denota una relación inversamente proporcional, indicando que a medida que una variable aumenta, la otra tiende a disminuir. Un coeficiente cercano a 0 sugiere que no hay una asociación lineal aparente entre las variables. La tabla de correlación es valiosa para identificar posibles patrones y tendencias entre variables, lo que permite una comprensión más profunda de la interacción entre los datos y guía el proceso de toma de decisiones en diversas áreas.

5. Utilice las variables categóricas, haga tablas de frecuencia, proporción, gráficas de barras o cualquier otra técnica que le permita explorar los datos.

```r
# Categorías de las variables cualitativas
categorias_cualitativas <- c("STDs.cervical.condylomatosis", "Schiller", "STDs.HPV", "IUD", "STDs.vulvo

tablas_frecuencias_cualitativas <- lapply(categorias_cualitativas, function(var) table(df[[var]]))

for (i in seq_along(tablas_frecuencias_cualitativas)) {
  cat("Tabla de frecuencias para la variable:", categorias_cualitativas[i], "\n")
  print(tablas_frecuencias_cualitativas[[i]])
  cat("\n")
}
```

```
## Tabla de frecuencias para la variable: STDs.cervical.condylomatosis
##
## 0.0
## 753
##
## Tabla de frecuencias para la variable: Schiller
##
##   0   1
## 784  74
##
## Tabla de frecuencias para la variable: STDs.HPV
##
## 0.0 1.0
## 751   2
##
## Tabla de frecuencias para la variable: IUD
##
## 0.0 1.0
## 658  83
##
## Tabla de frecuencias para la variable: STDs.vulvo.perineal.condylomatosis
##
## 0.0 1.0
## 710  43
```

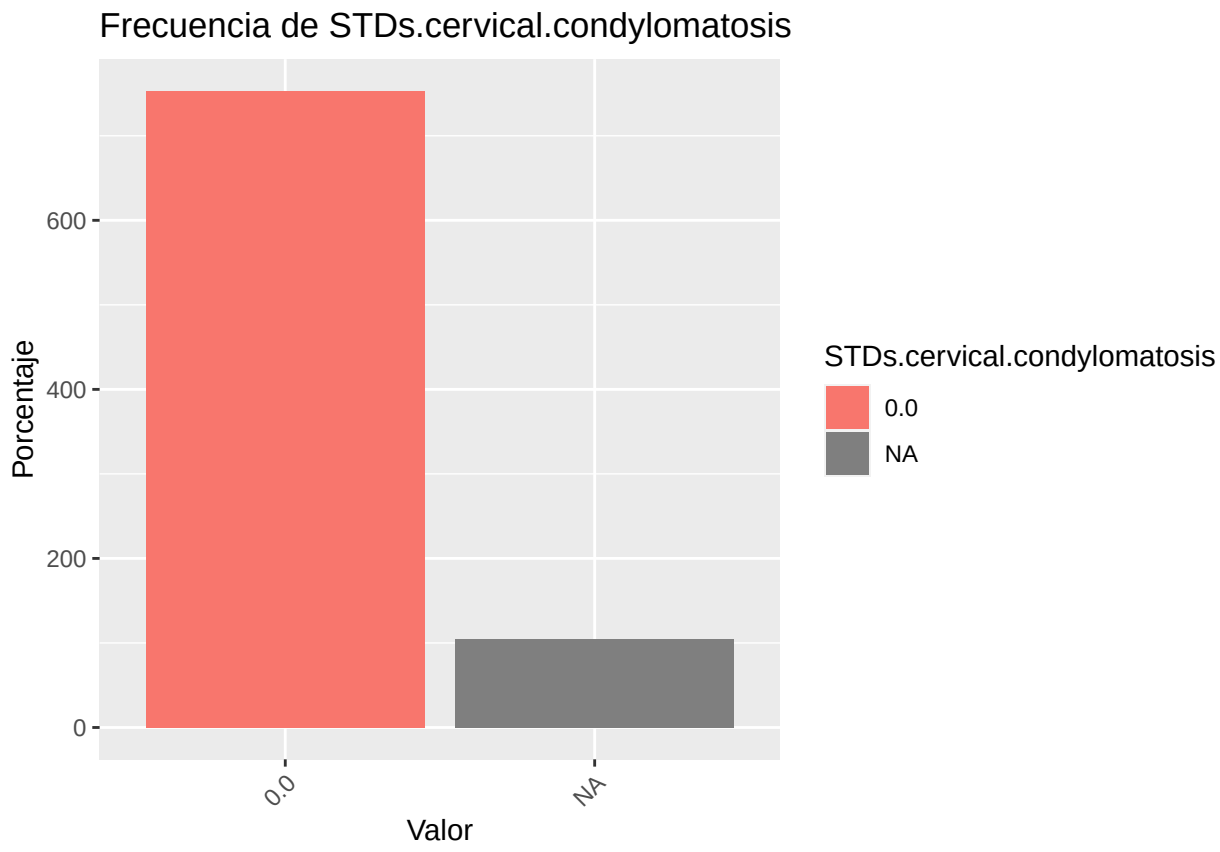
```
##
## Tabla de frecuencias para la variable: STDs
##
## 0.0 1.0
## 674  79
##
## Tabla de frecuencias para la variable: STDs.syphilis
##
## 0.0 1.0
## 735  18
##
## Tabla de frecuencias para la variable: Dx.HPV
##
##   0   1
## 840  18
##
## Tabla de frecuencias para la variable: Dx.Cancer
##
##   0   1
## 840  18
##
## Tabla de frecuencias para la variable: STDs.HIV
##
## 0.0 1.0
## 735  18
##
## Tabla de frecuencias para la variable: STDs.molluscum.contagiosum
##
## 0.0 1.0
## 752   1
##
## Tabla de frecuencias para la variable: STDs.condylomatosis
##
## 0.0 1.0
## 709  44
##
## Tabla de frecuencias para la variable: Dx.CIN
##
##   0   1
## 849   9
##
## Tabla de frecuencias para la variable: Hinselmann
##
##   0   1
## 823  35
##
## Tabla de frecuencias para la variable: STDs.Hepatitis.B
##
## 0.0 1.0
## 752   1
##
## Tabla de frecuencias para la variable: Hormonal.Contraceptives
##
## 0.0 1.0
```

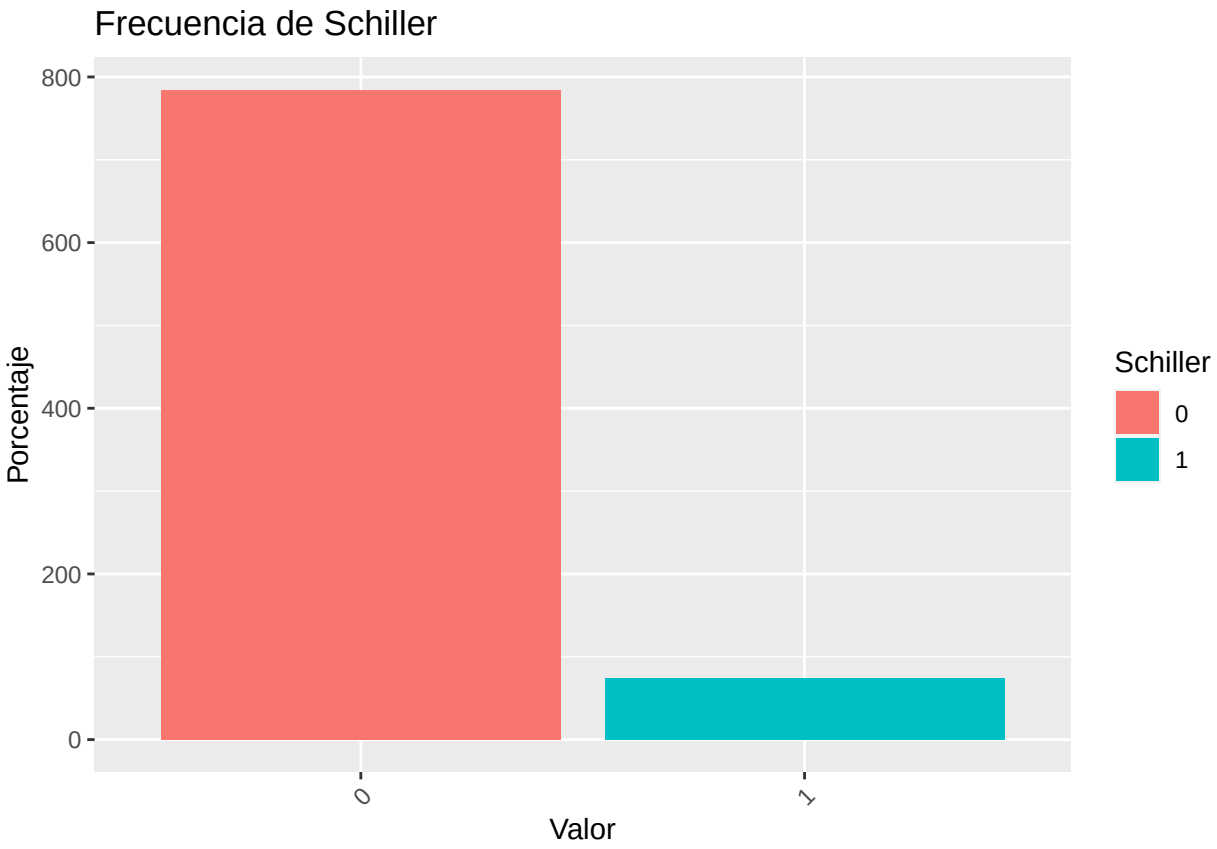
```
## 269 481
##
## Tabla de frecuencias para la variable: Biopsy
##
##   0   1
## 803  55
##
## Tabla de frecuencias para la variable: Citology
##
##   0   1
## 814  44
##
## Tabla de frecuencias para la variable: STDs.genital.herpes
##
## 0.0 1.0
## 752   1
##
## Tabla de frecuencias para la variable: STDs.vaginal.condylomatosis
##
## 0.0 1.0
## 749   4
##
## Tabla de frecuencias para la variable: Dx
##
##   0   1
## 834  24
##
## Tabla de frecuencias para la variable: STDs.AIDS
##
## 0.0
## 753
##
## Tabla de frecuencias para la variable: Smokes
##
## 0.0 1.0
## 722 123
```

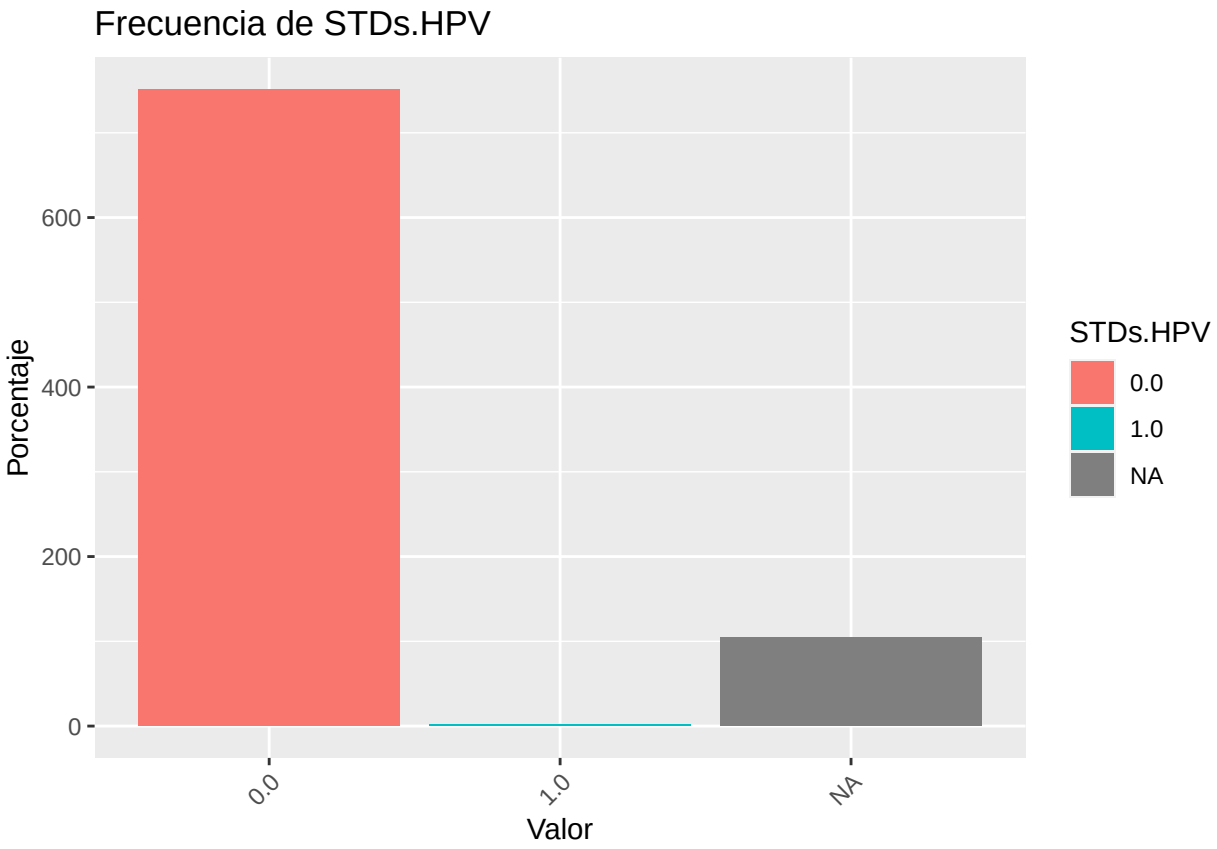
Para variables categóricas, la tabla de frecuencias nos proporciona información sobre la distribución de las categorías y nos permite identificar las categorías más comunes y menos comunes. Además, nos permite identificar si hay categorías que dominan la distribución o si existe un equilibrio en la frecuencia entre las categorías. Como se puede observar varias variables nos dan datos interesantes que nos pueden servir despues si queremos predecir.

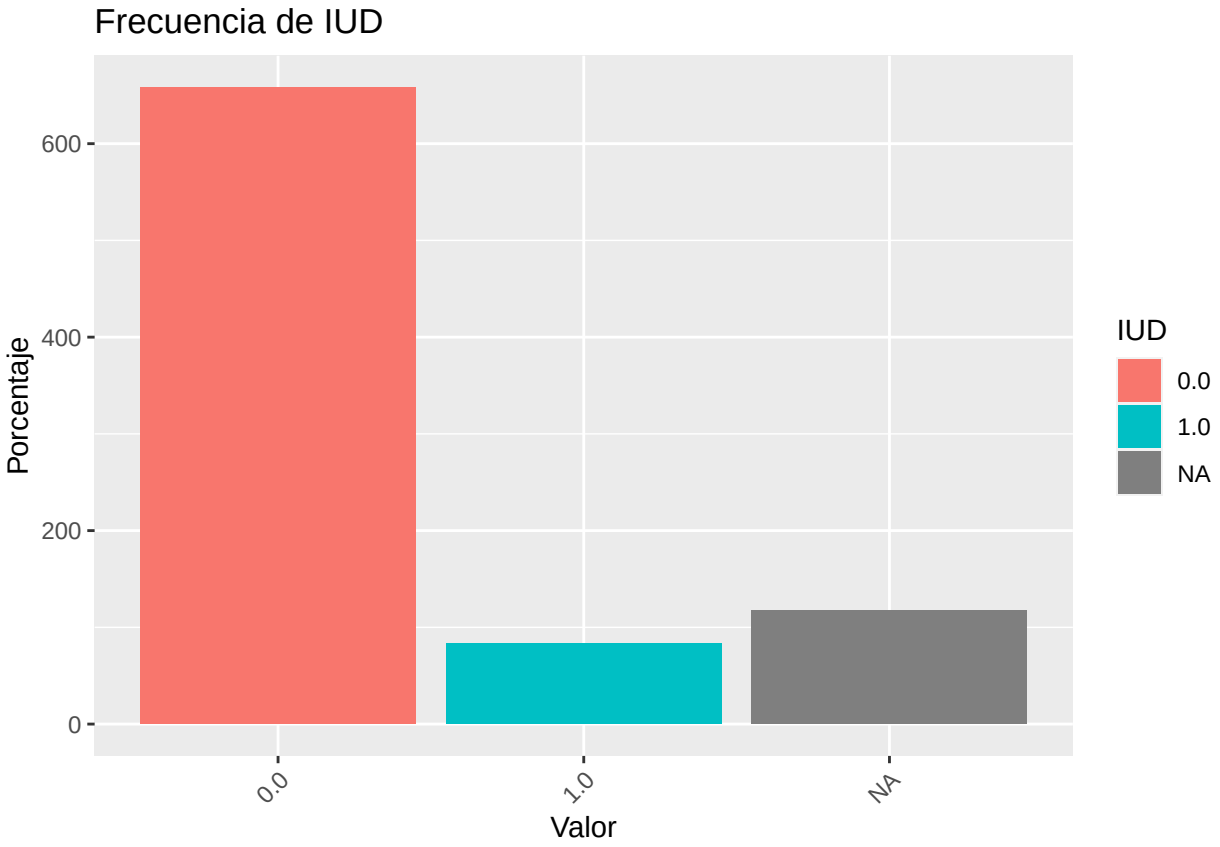
```
# Crear una función para generar gráficas de barras para las variables categóricas
graficar_barras <- function(data, variable) {
  # Calcular la tabla de frecuencias
  tabla_frecuencia <- prop.table(table(data[[variable]])) * 100

  # Ordenar las categorías de mayor a menor frecuencia
  orden <- names(sort(-tabla_frecuencia))
  data[[variable]] <- factor(data[[variable]], levels = orden)

  # Crear la gráfica de barras
  p <- ggplot(data = data, aes(x = .data[[variable]], fill = .data[[variable]])) +
```

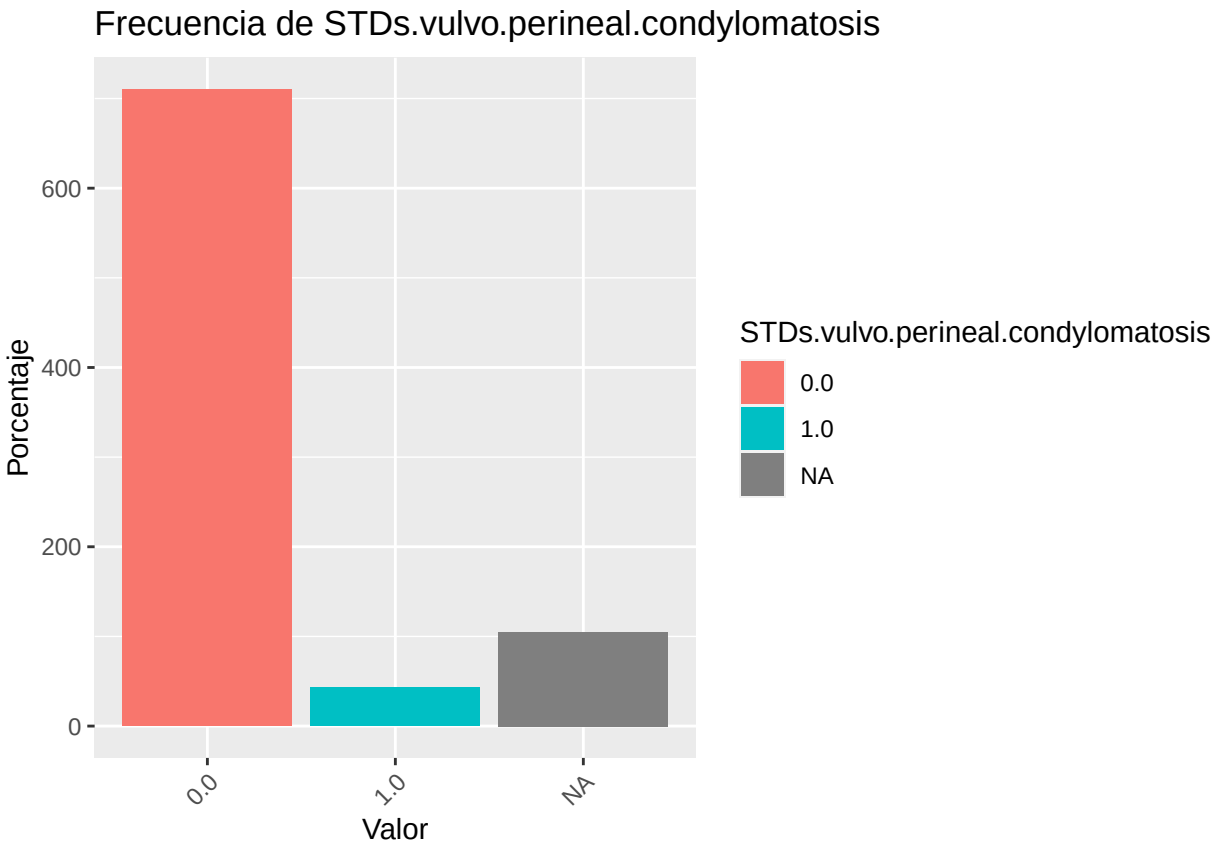
```r
    geom_bar() +
    labs(title = paste("Frecuencia de", variable), x = "Valor", y = "Porcentaje") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))

  return(p)
}

# Crear una ventana gráfica para mostrar las gráficas de barras
par(mfrow = c(6, 4))  # Configuramos el diseño para mostrar varias gráficas en una sola ventana

# Generar las gráficas de barras para las variables categóricas cualitativas
for (variable in categorias_cualitativas) {
  grafica <- graficar_barras(df, variable)
  print(grafica)
}
```

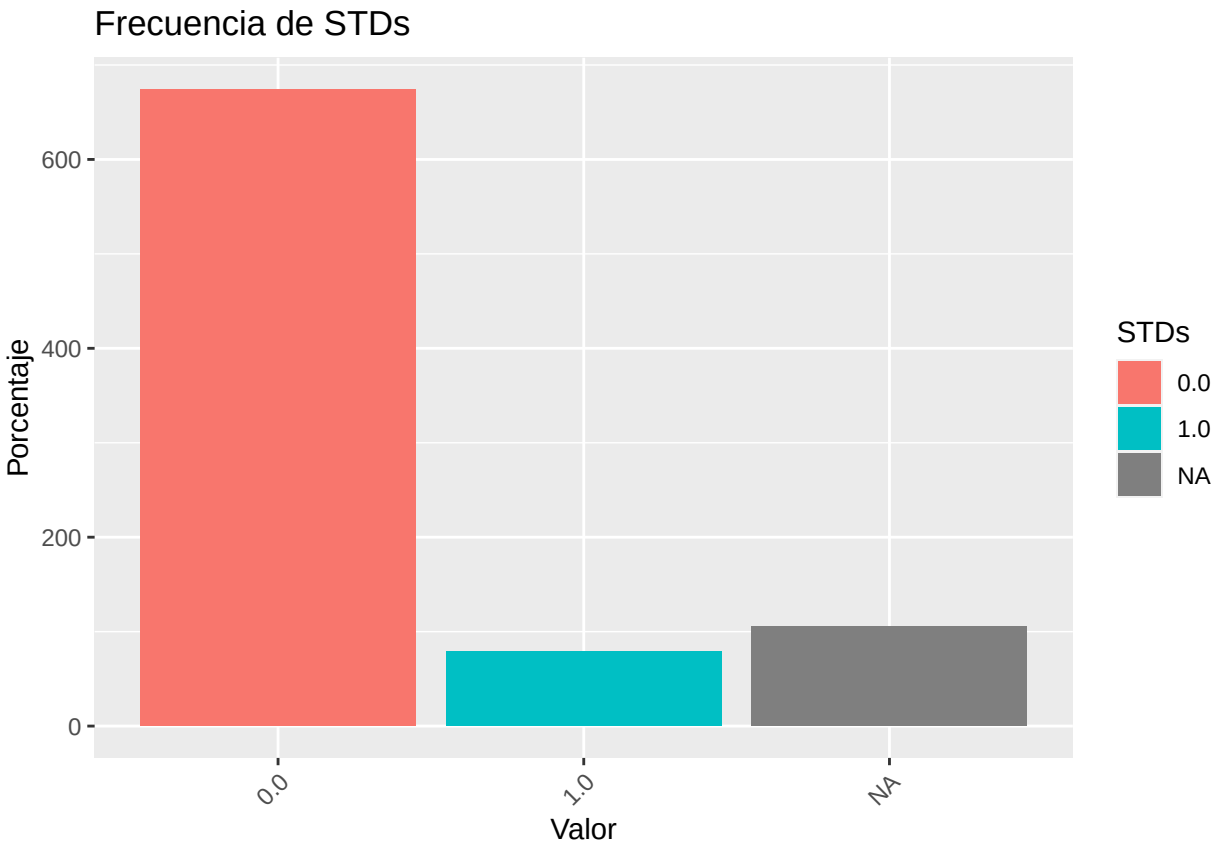



Frecuencia de Schiller
800
600
400
200
0
Porcentaje
0
1
Valor
Schiller
0
1

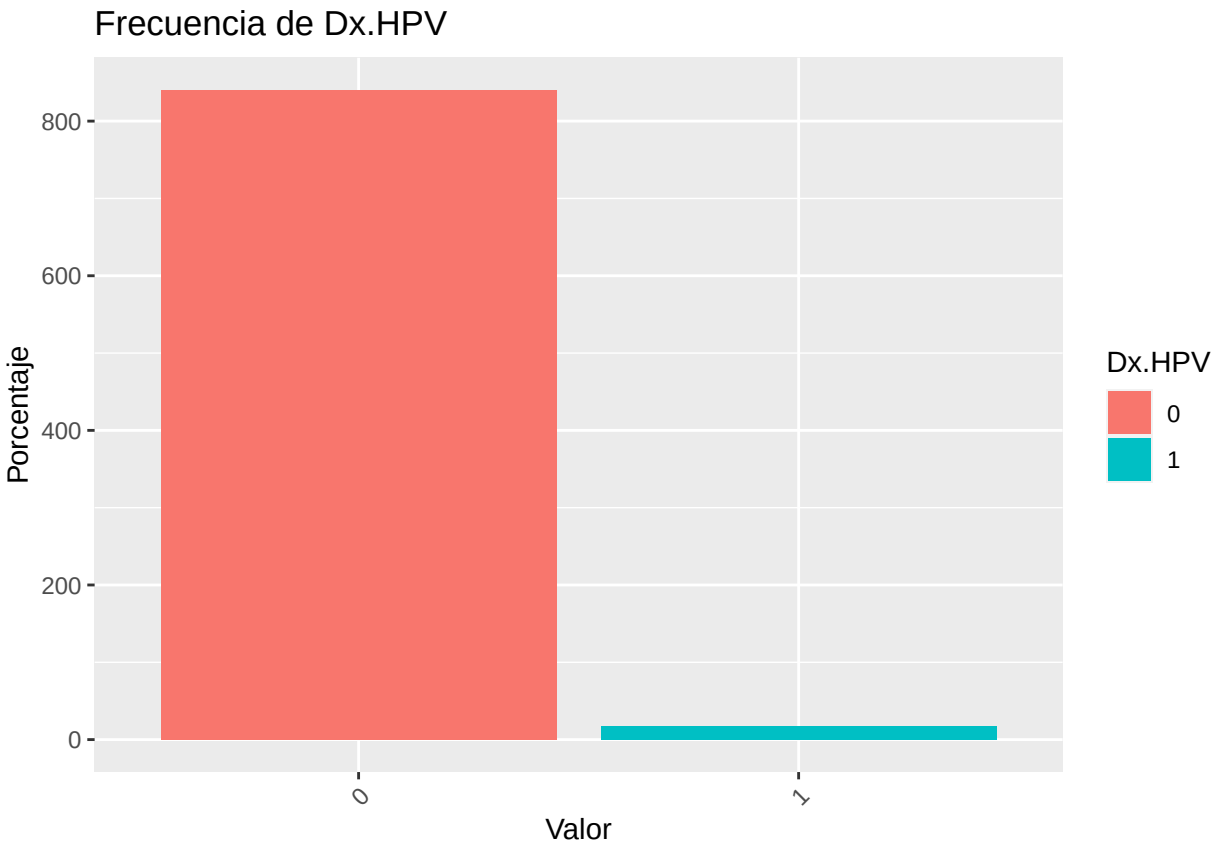

Frecuencia de STDs.HPV
Porcentaje
600
400
200
0
0.0
1.0
NA
Valor
STDs.HPV
0.0
1.0
NA

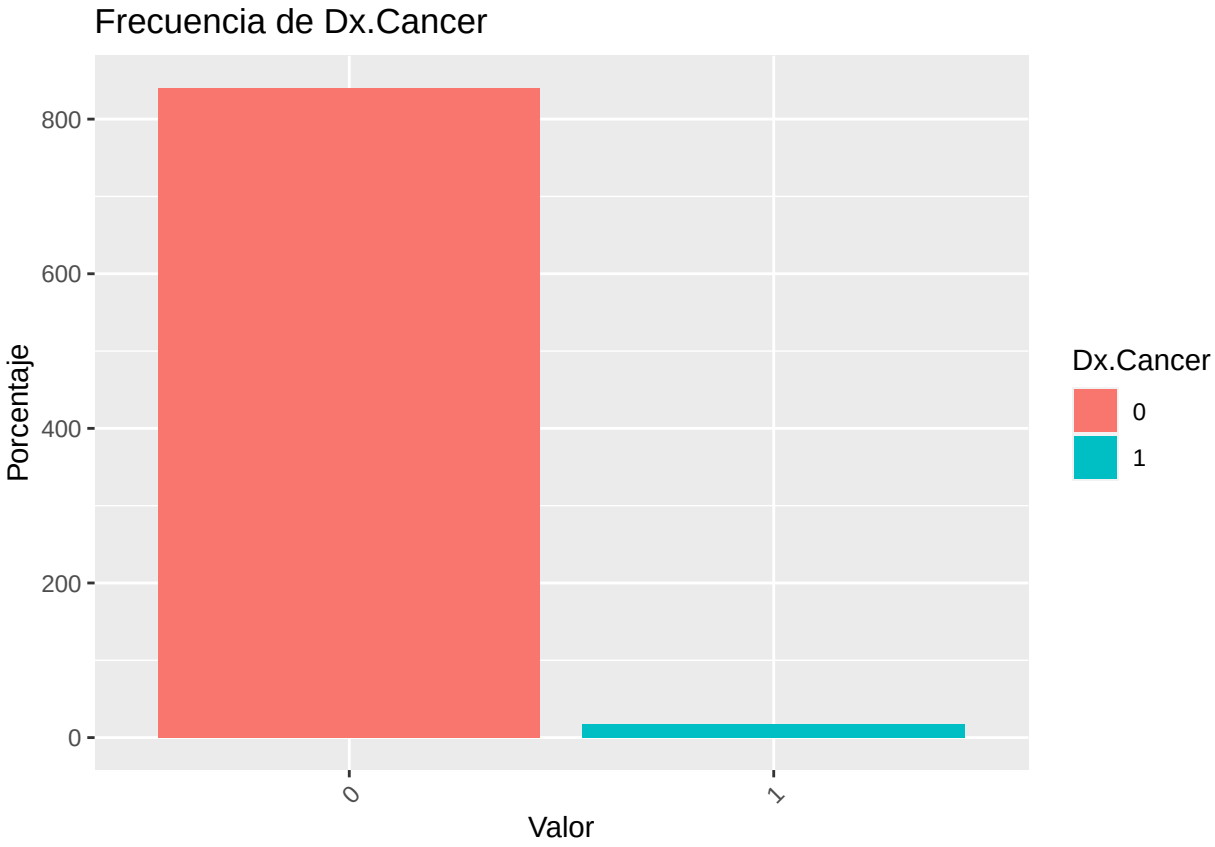

Frecuencia de IUD
600
400
200
0
Porcentaje
0.0
1.0
NA
Valor
IUD
0.0
1.0
NA

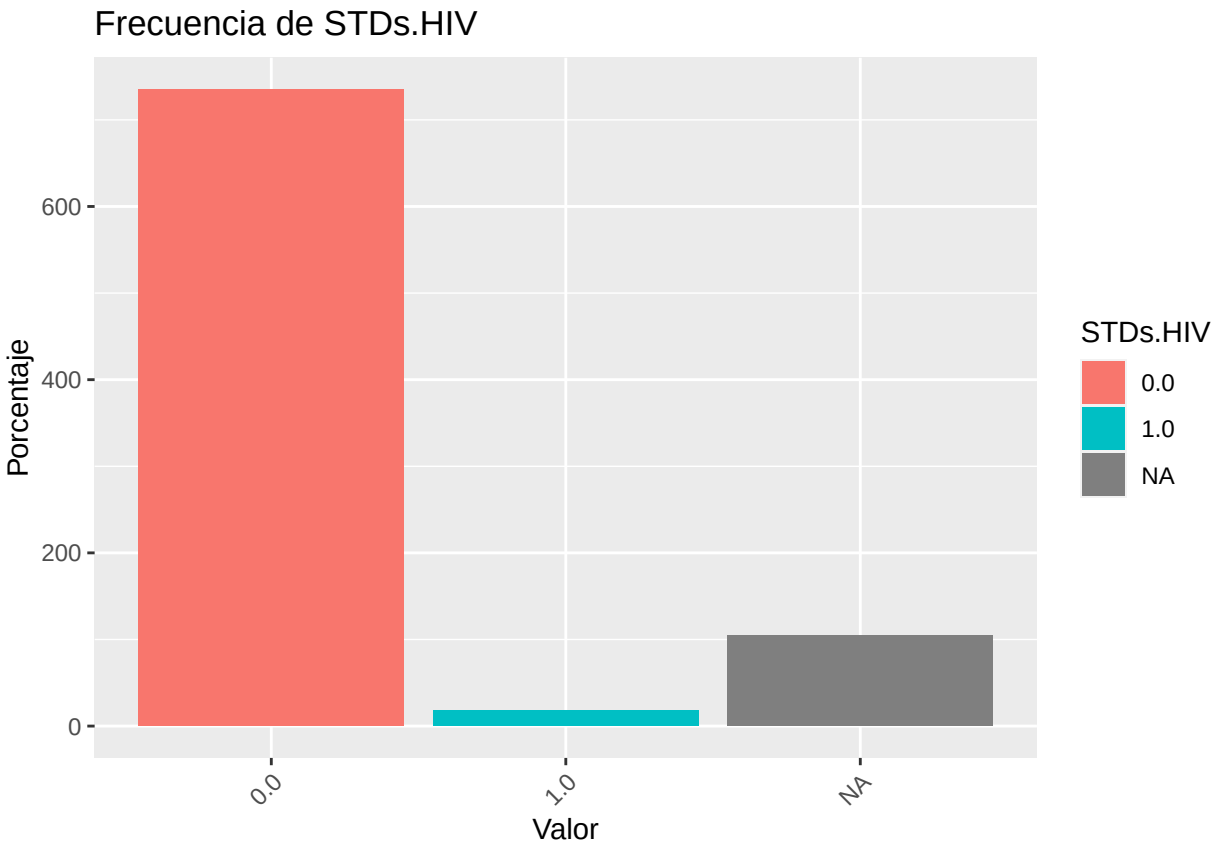

Frecuencia de STDs.vulvo.perineal.condylomatosis
Porcentaje
600
400
200
0
0.0
1.0
NA
Valor
STDs.vulvo.perineal.condylomatosis
0.0
1.0
NA

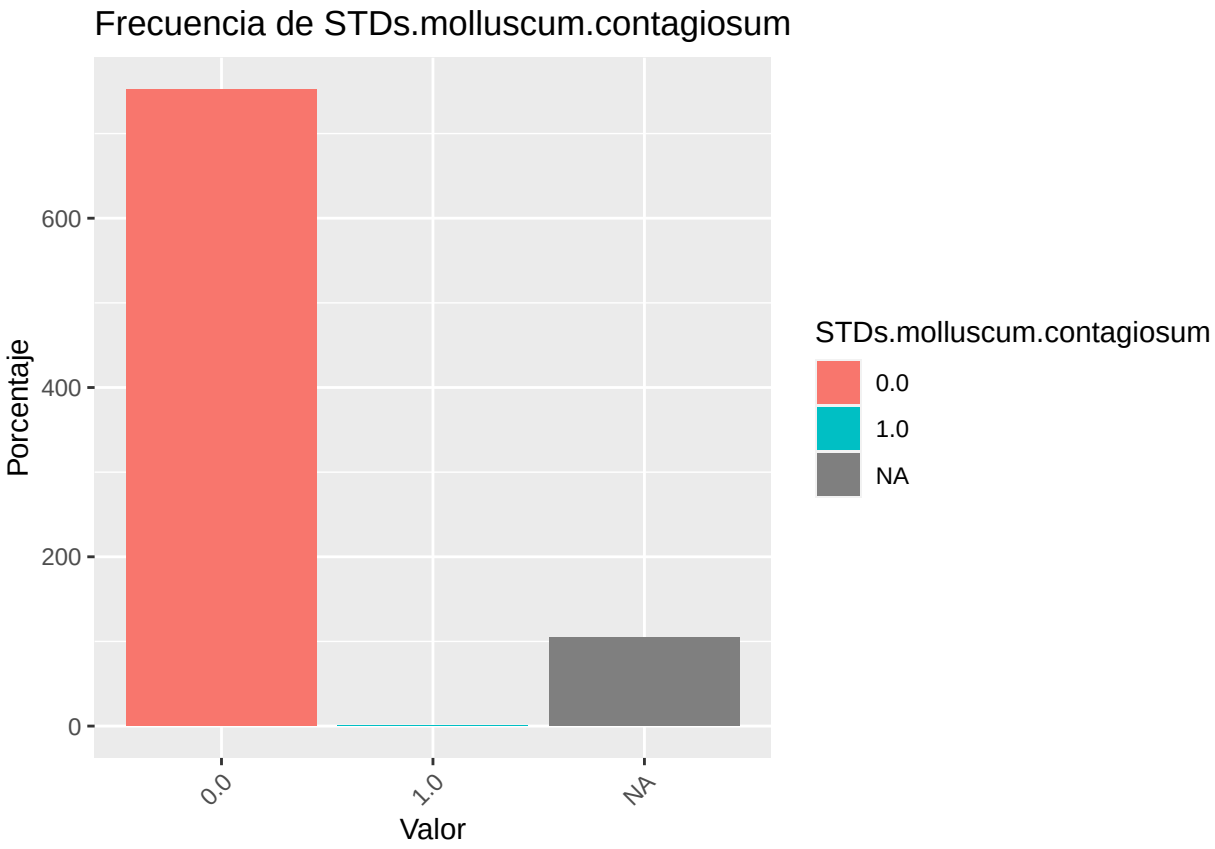

Frecuencia de STDs
Porcentaje
600
400
200
0
0.0
1.0
NA
Valor
STDs
0.0
1.0
NA

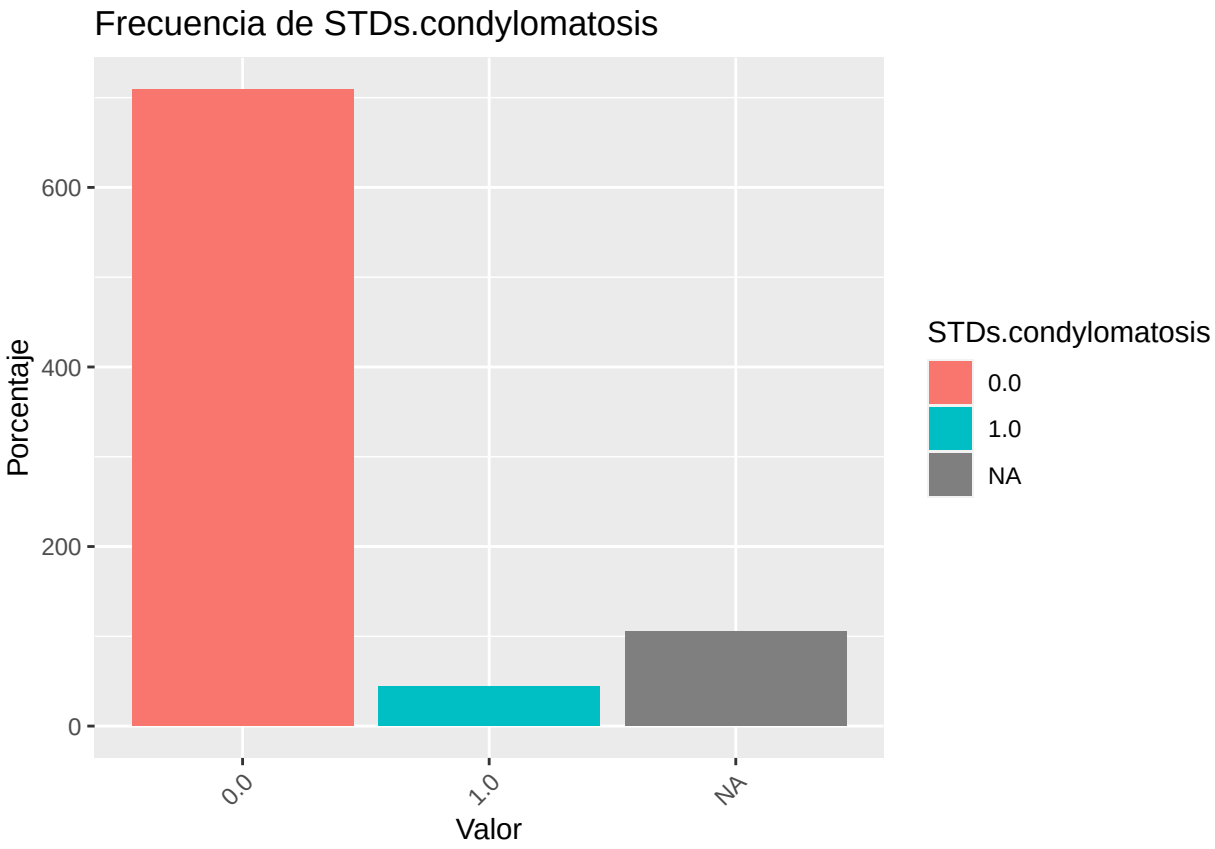

Frecuencia de STDs.syphilis
Porcentaje
600
400
200
0
0.0
1.0
NA
Valor
STDs.syphilis
0.0
1.0
NA

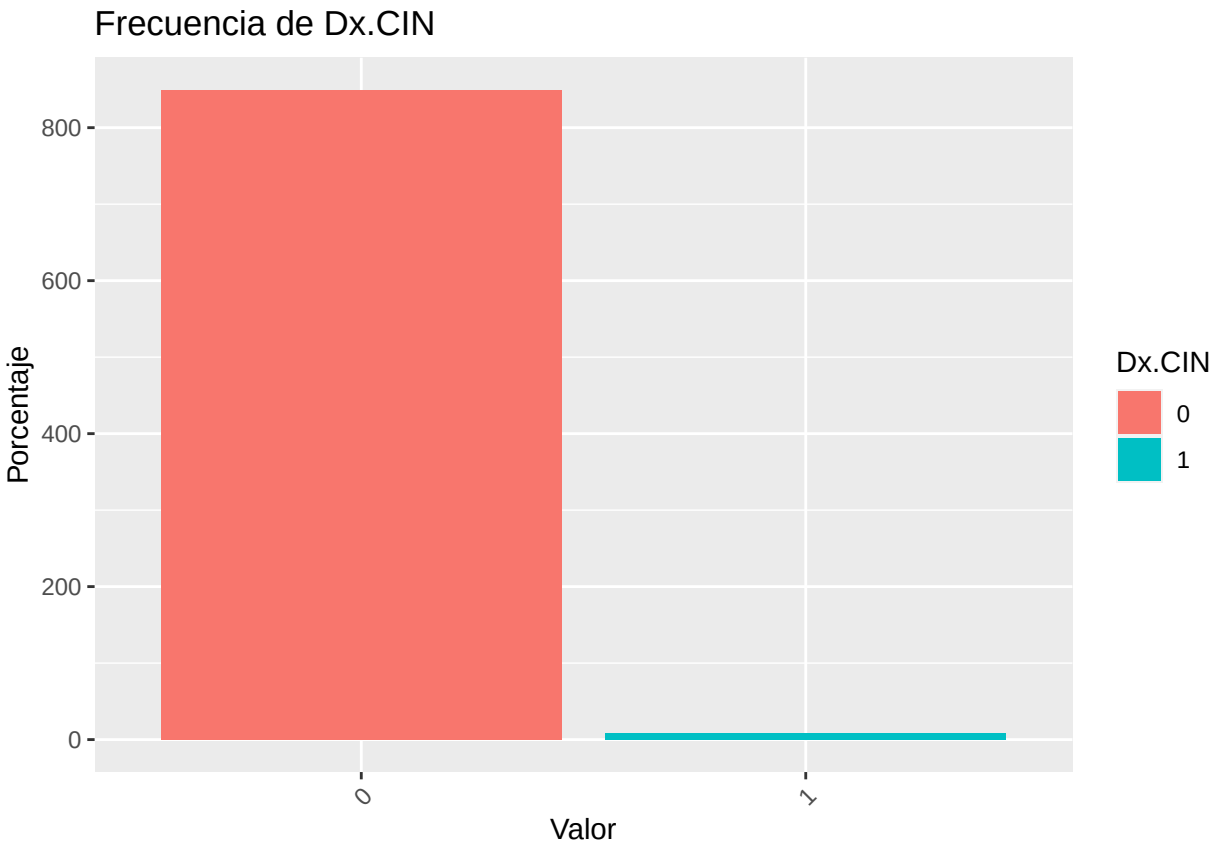

Frecuencia de Dx.HPV
800
600
400
200
0
Porcentaje
0
1
Valor
Dx.HPV
0
1

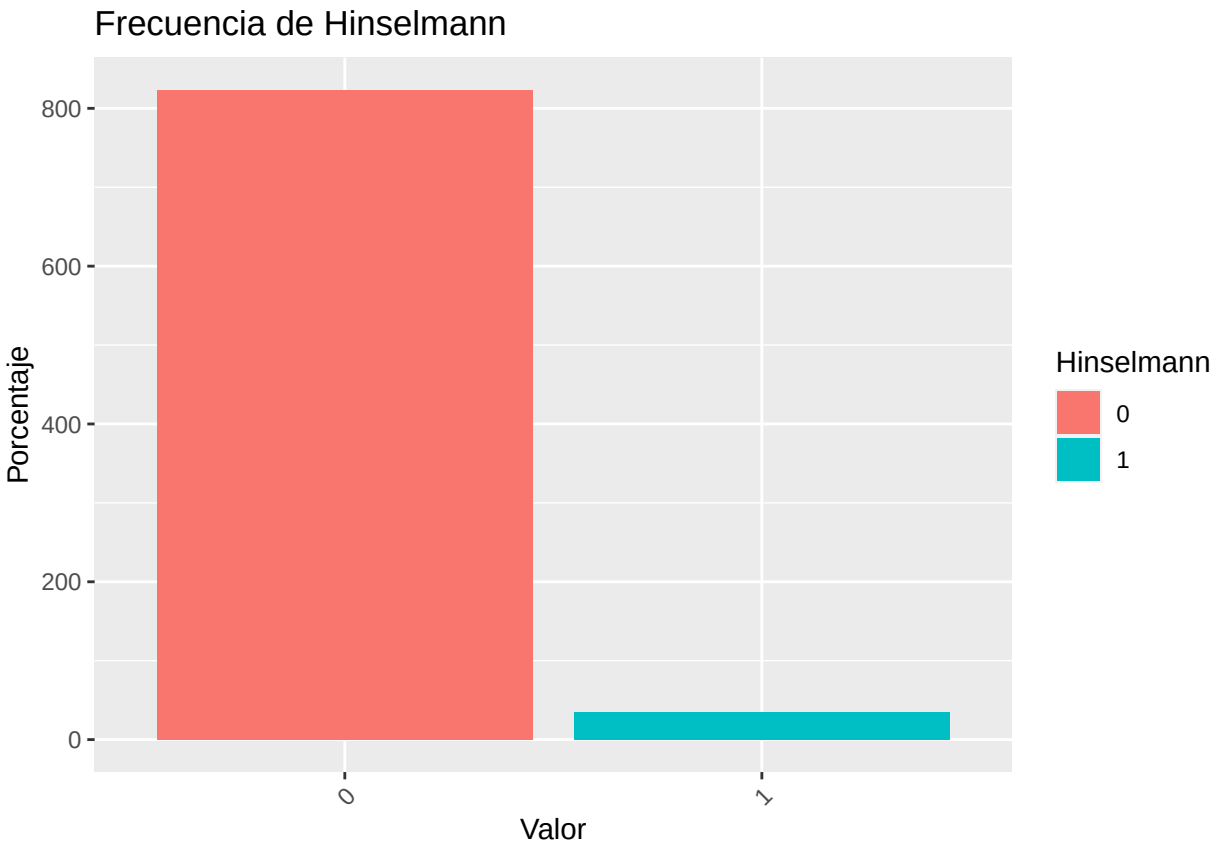

Frecuencia de Dx.Cancer
800
600
400
200
0
Porcentaje
0
1
Valor
Dx.Cancer
0
1

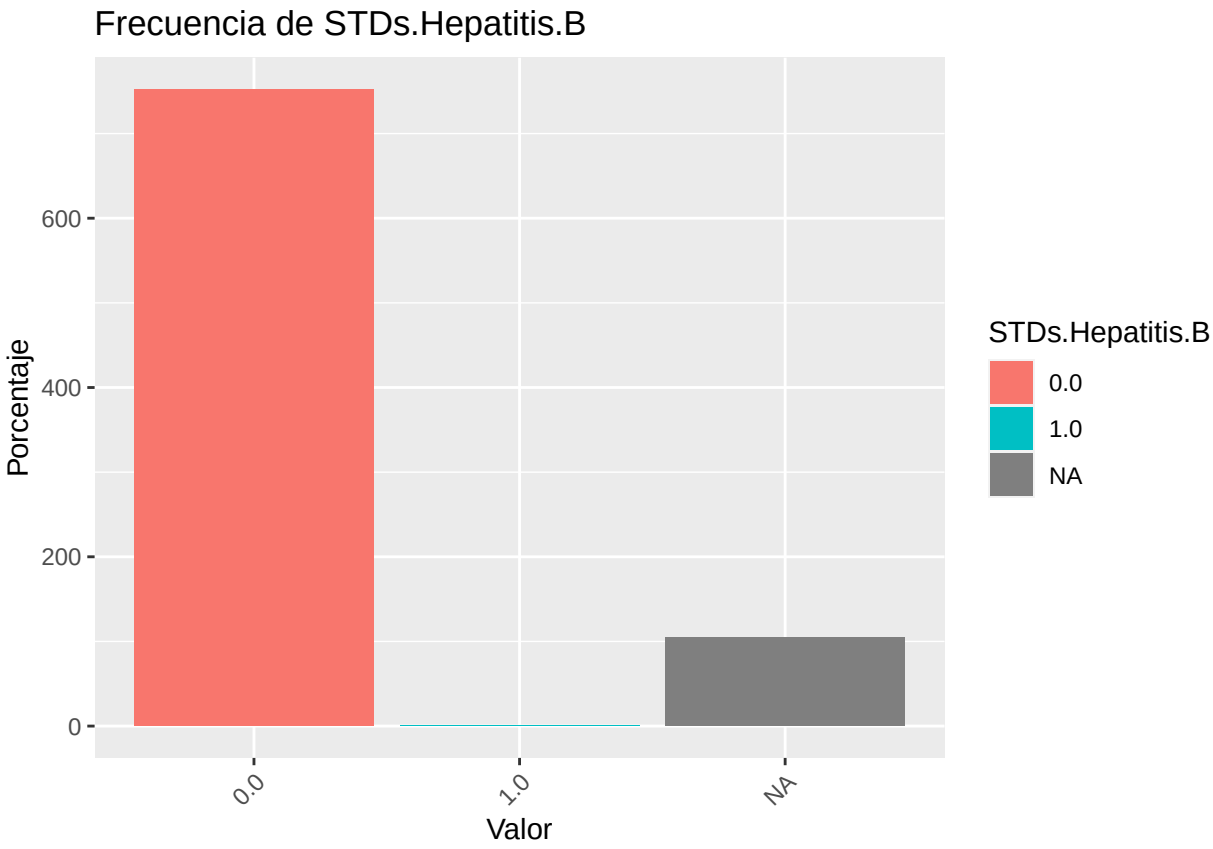

Frecuencia de STDs.HIV
Porcentaje
600
400
200
0
0.0
1.0
NA
Valor
STDs.HIV
0.0
1.0
NA

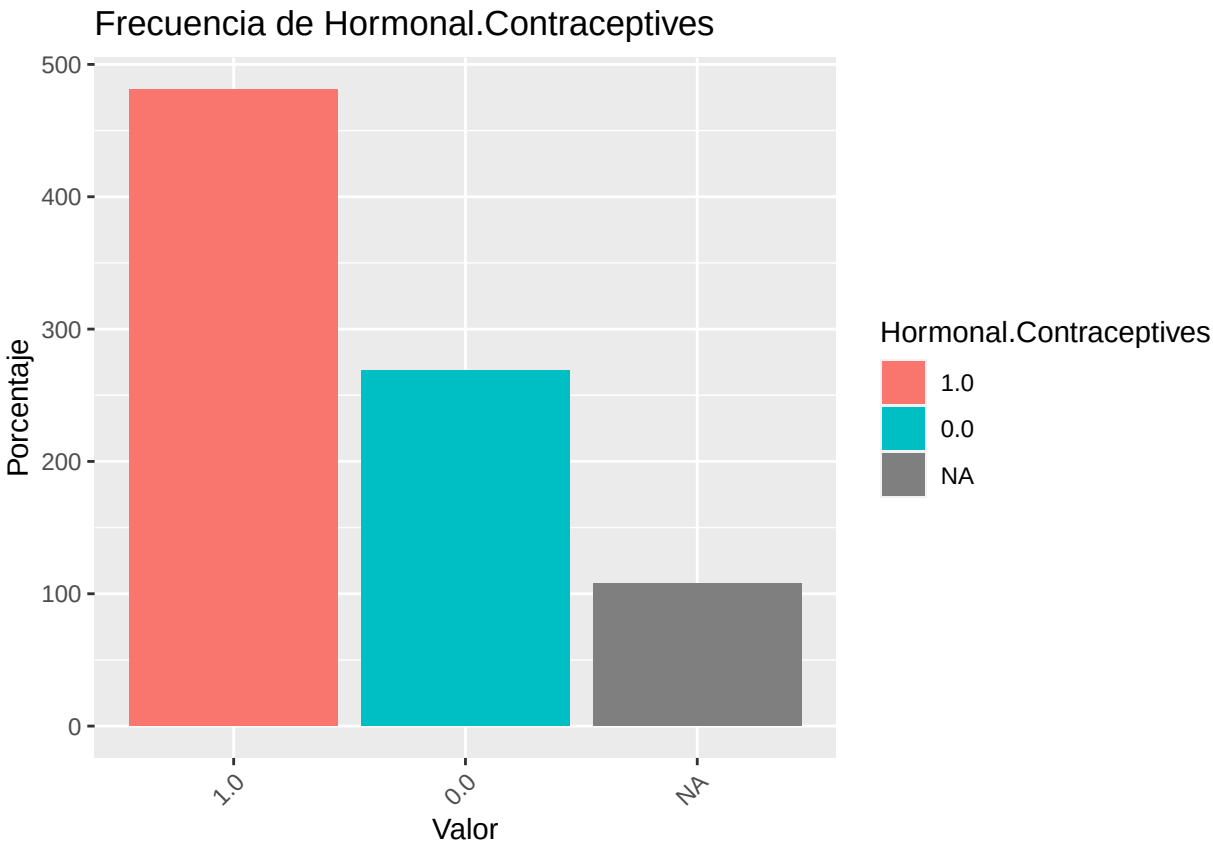

Frecuencia de STDs.molluscum.contagiosum
Porcentaje
600
400
200
0
0.0
1.0
NA
Valor
STDs.molluscum.contagiosum
0.0
1.0
NA

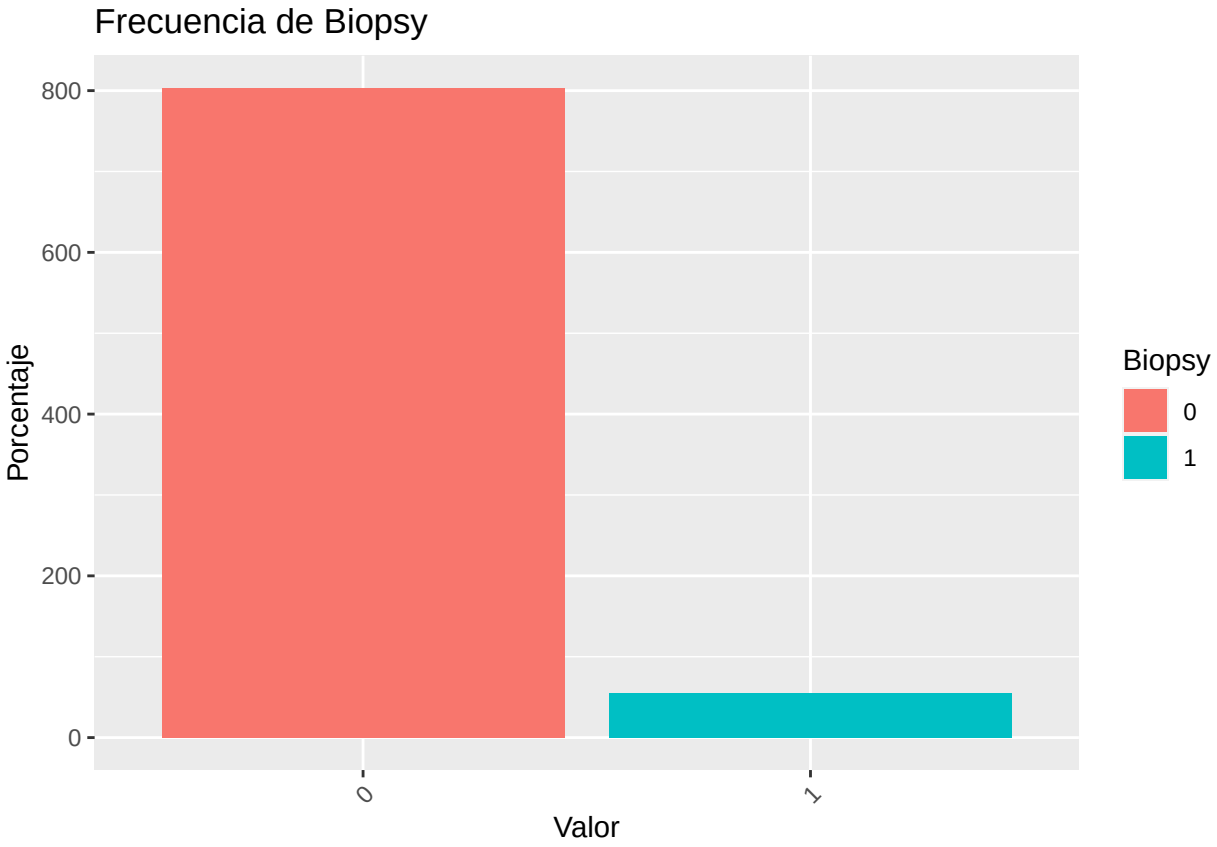

Frecuencia de STDs.condylomatosis
Porcentaje
600
400
200
0
0.0
1.0
NA
Valor
STDs.condylomatosis
0.0
1.0
NA

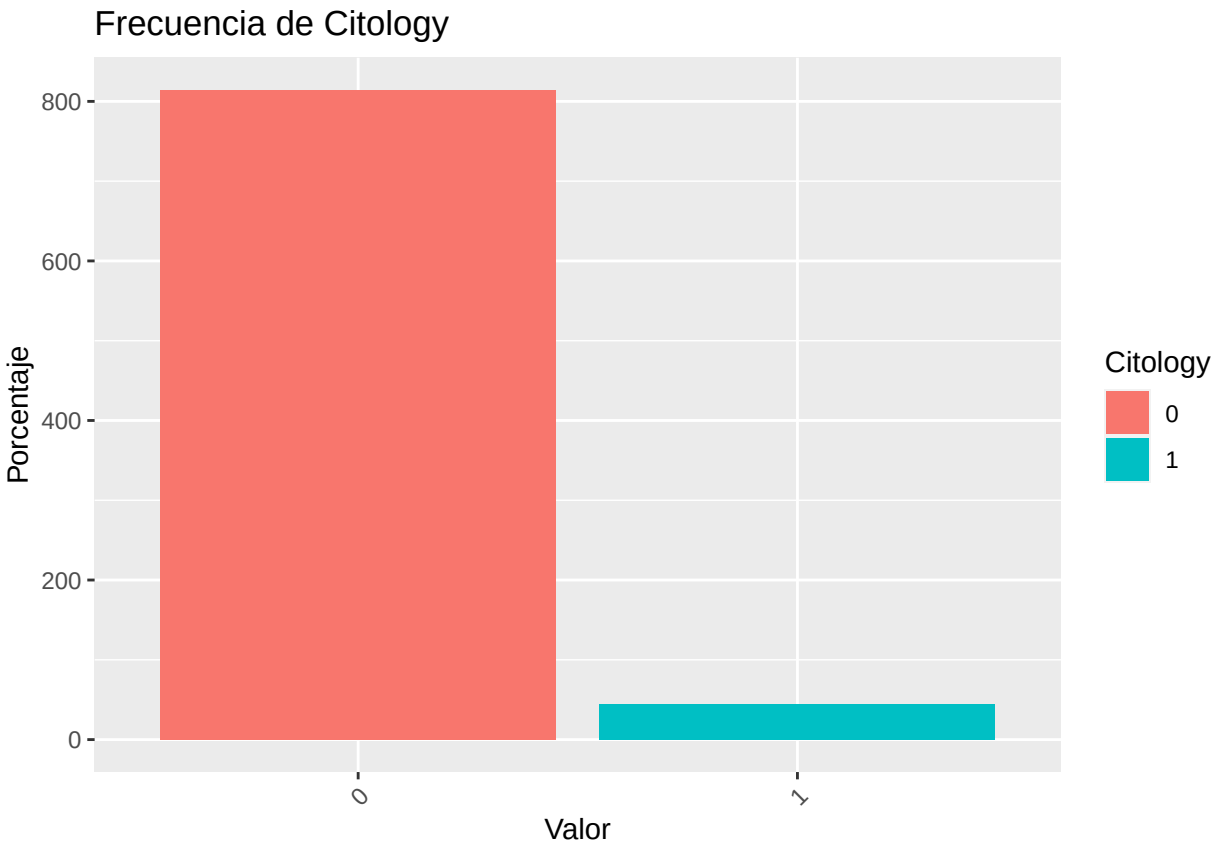

Frecuencia de Dx.CIN
800
600
400
200
0
Porcentaje
0
1
Valor
Dx.CIN
0
1

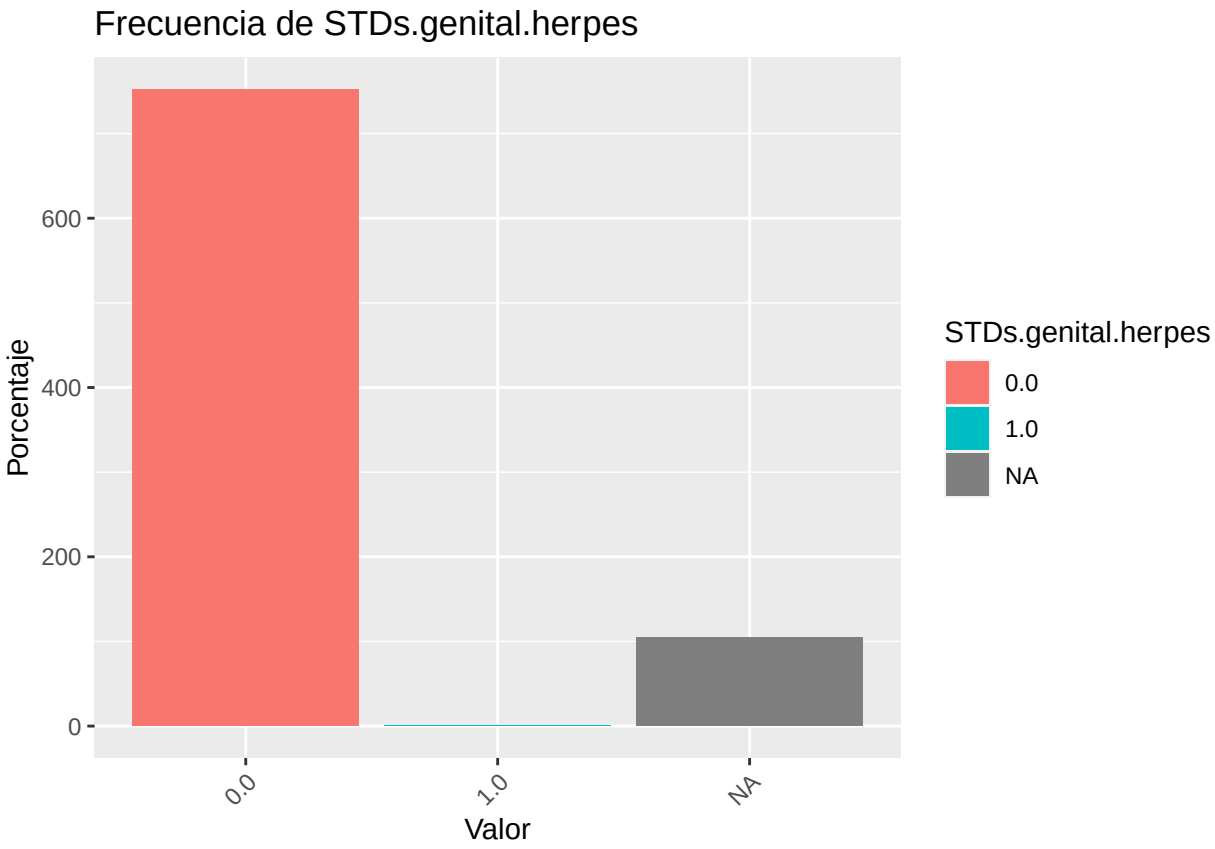

Frecuencia de Hinselmann
800
600
400
200
0
Porcentaje
0
1
Valor
Hinselmann
0
1

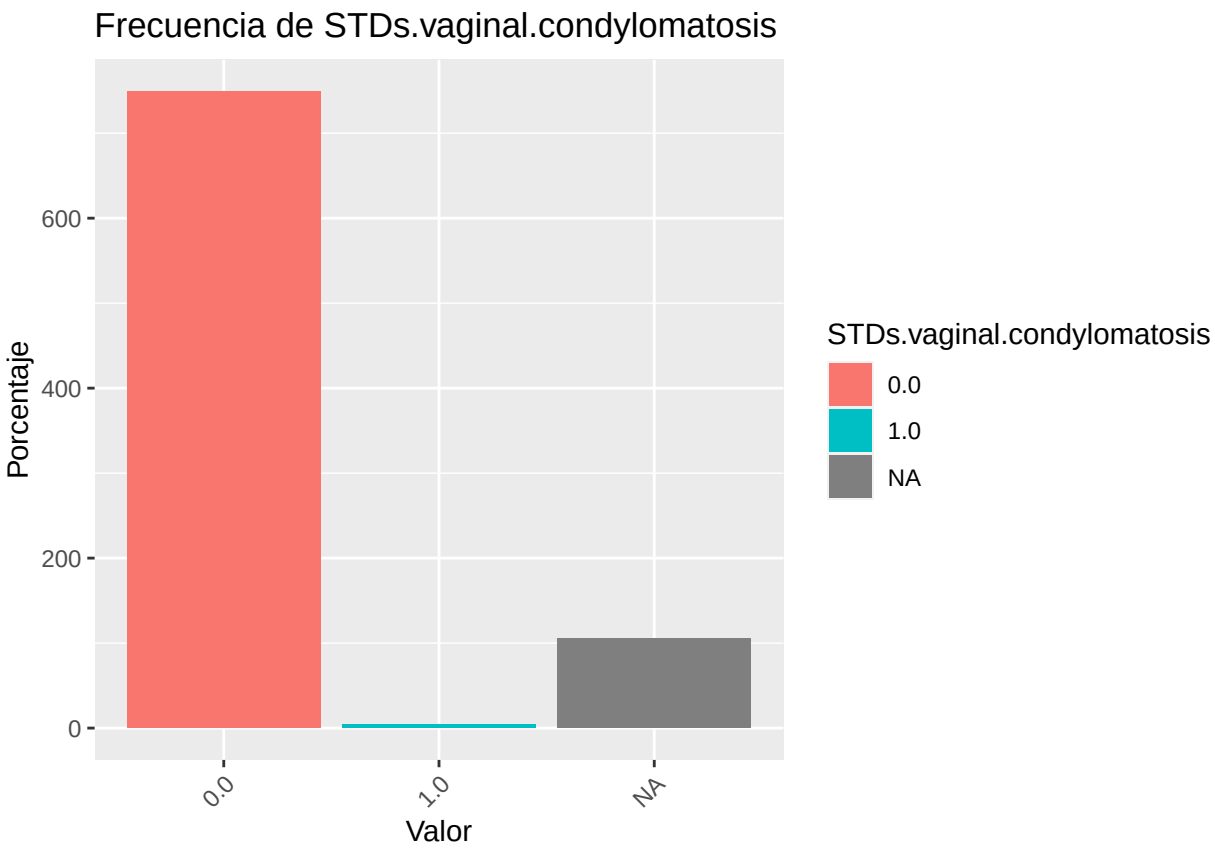

Frecuencia de STDs.Hepatitis.B
Porcentaje
600
400
200
0
0.0
1.0
NA
Valor
STDs.Hepatitis.B
0.0
1.0
NA

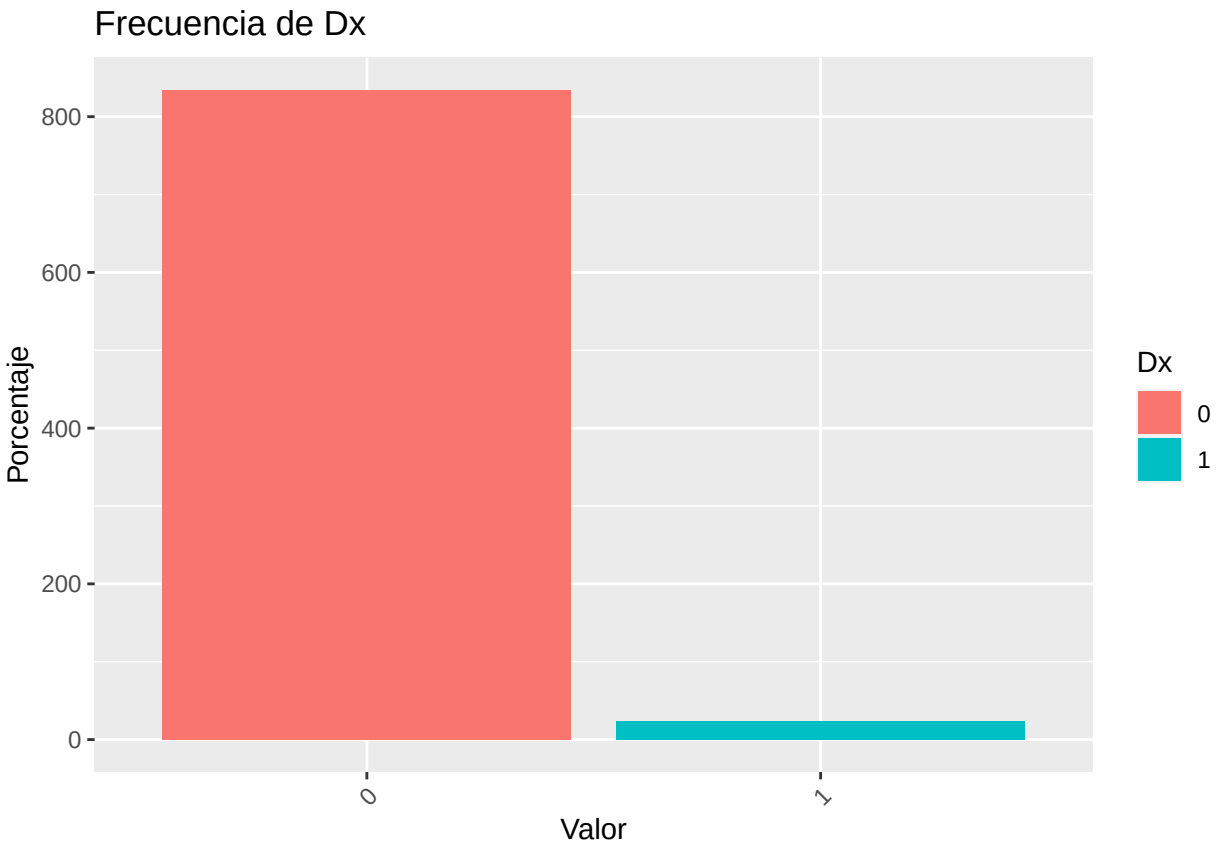

Frecuencia de Hormonal.Contraceptives
500
400
300
200
100
0
Porcentaje
1.0
0.0
NA
Valor
Hormonal.Contraceptives
1.0
0.0
NA

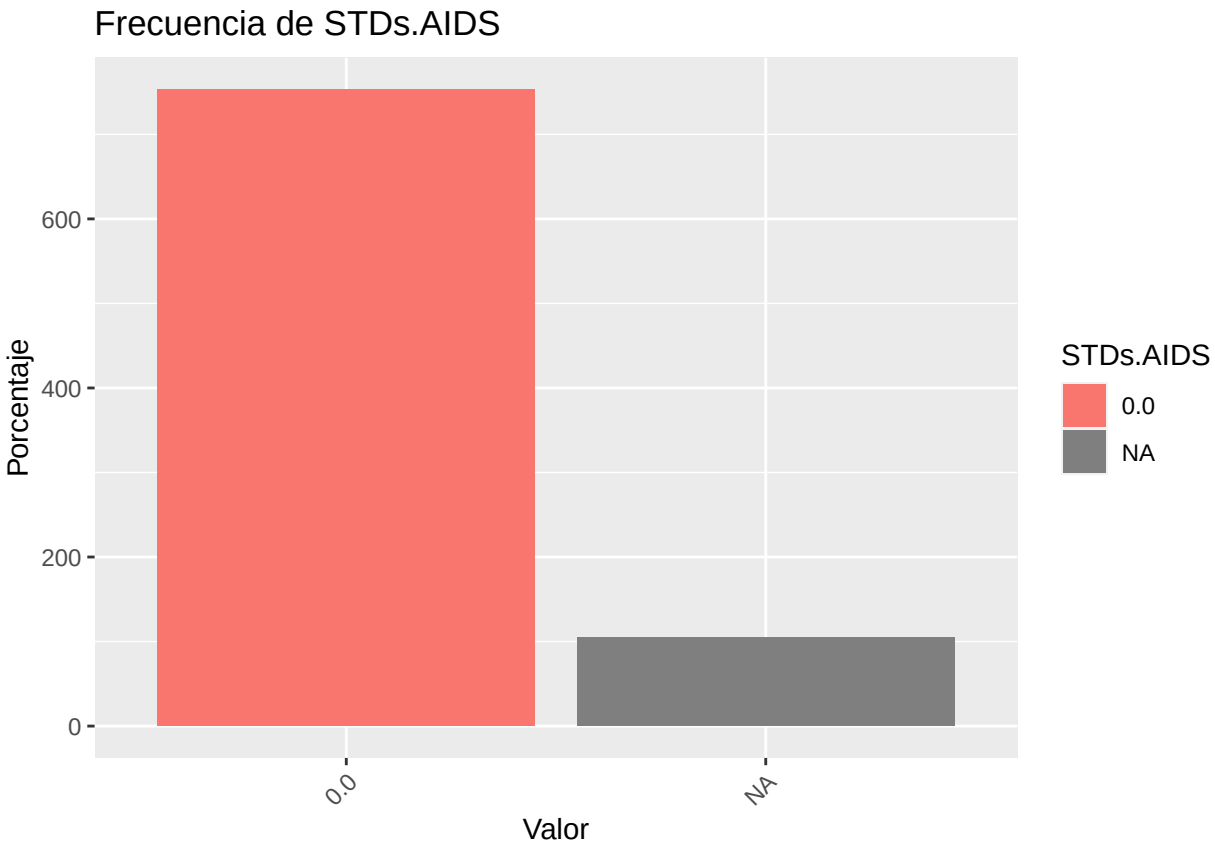

Frecuencia de Biopsy
800
600
400
200
0
Porcentaje
0
1
Valor
Biopsy
0
1

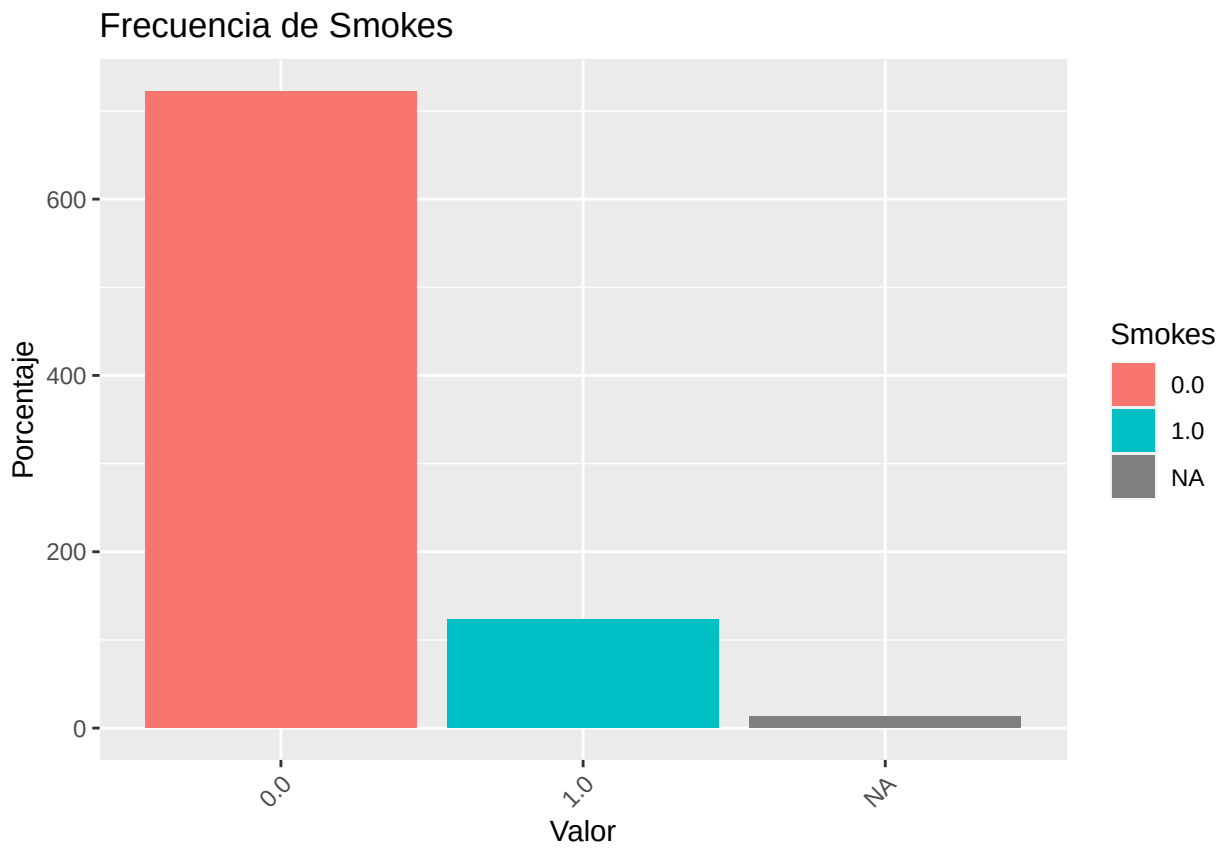

Frecuencia de Citology
800
600
400
200
0
Porcentaje
0
1
Valor
Citology
0
1


Frecuencia de STDs.genital.herpes
Porcentaje
600
400
200
0
0.0
1.0
NA
Valor
STDs.genital.herpes
0.0
1.0
NA


Frecuencia de STDs.vaginal.condylomatosis
Porcentaje
600
400
200
0
0.0
1.0
NA
Valor
STDs.vaginal.condylomatosis
0.0
1.0
NA

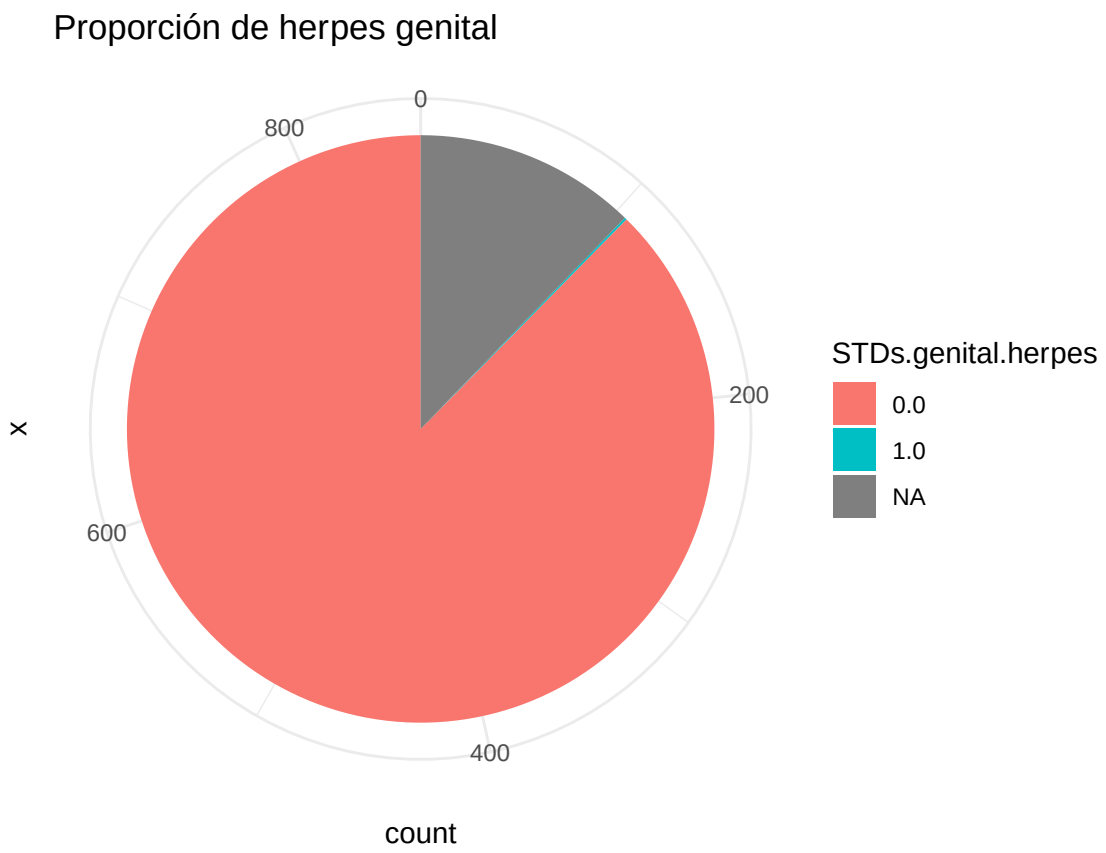

Frecuencia de Dx
800
600
400
200
0
Porcentaje
0
1
Valor
Dx
0
1


Frecuencia de STDs.AIDS
Porcentaje
600
400
200
0
0.0
NA
Valor
STDs.AIDS
0.0
NA



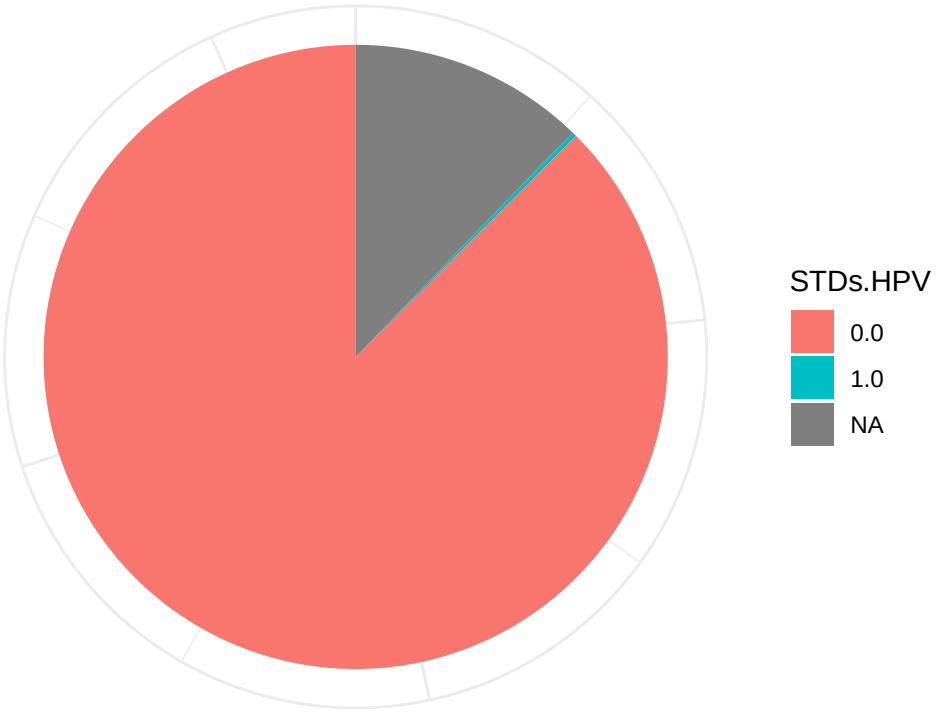
En los graficos se puede obervar cada una de las variables que lo que mas tiene cada uno son 0, con un poco de 1 y muy pocos NA esto nos ayuda a observar de mejoer manera que es lo que tiene cada una de las variables y averiguar lo que se va hacer a posterior.

```
ggplot(df, aes(x = "", fill = `STDs.genital.herpes`)) +
  geom_bar(width = 1) +
  coord_polar(theta = "y") +
  labs(title = "Proporción de herpes genital", fill = "STDs.genital.herpes") +
  theme_minimal()
```



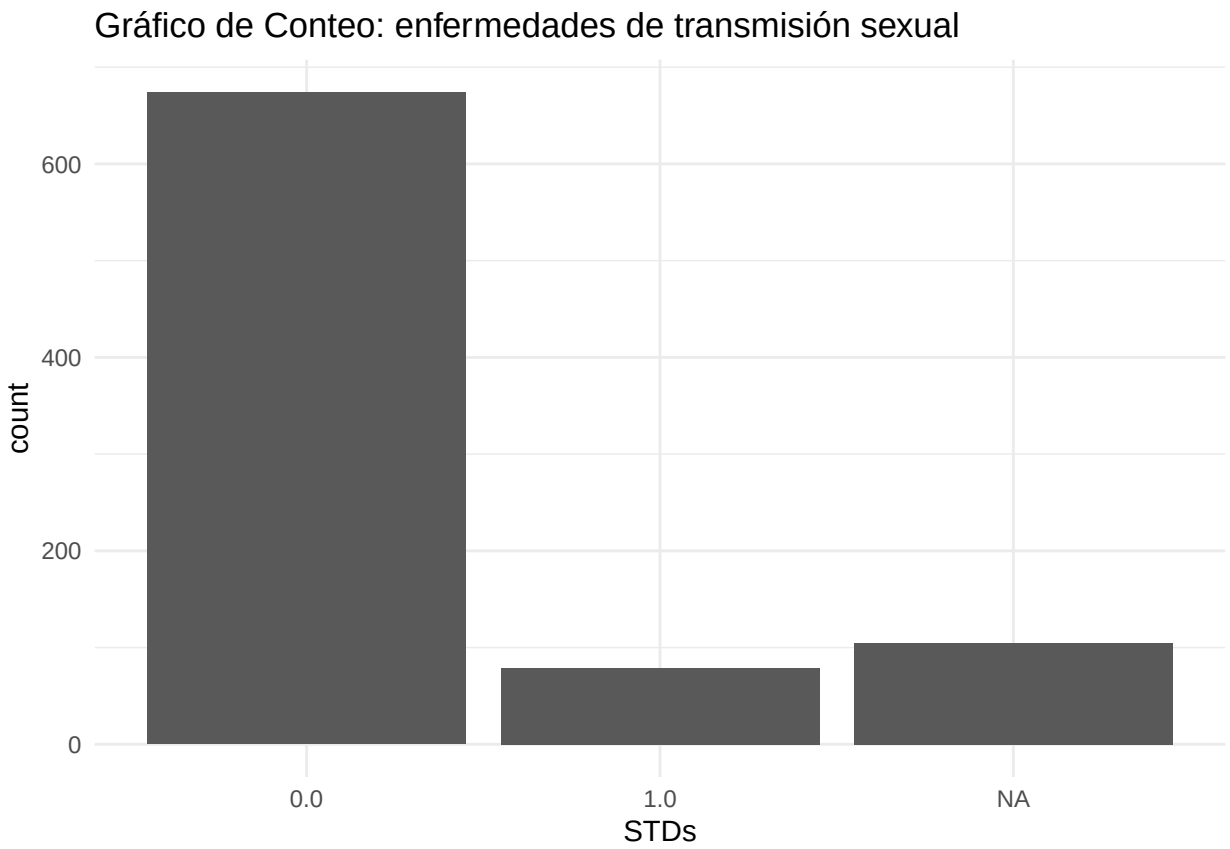
Vemos que la proporcion de herpes genital en su mayoria esta lleno de 0.0 y tambien unos cuantos NA y muy pocos 1.0 lo que significa que tenemos una cantidad de gente que tiene relacion con esta enfermedad o que prefiere no decirlo.

```
data_categoricas <- df[, c(categorias_cualitativas)]

ggplot(data_categoricas, aes(x = "", fill = `STDs.HPV`)) +
  geom_bar(width = 1) +
  coord_polar("y", start = 0) +
  labs(title = "Gráfico de Torta: Proporción de Personas con Papiloma Humano", x = NULL, y = NULL) +
  theme_minimal() +
  theme(axis.text.x = element_blank())
```



Este grafico de pie es similar al anterior la gran mayoria son 0.0 y bastantes NA y uy pocos 1.0

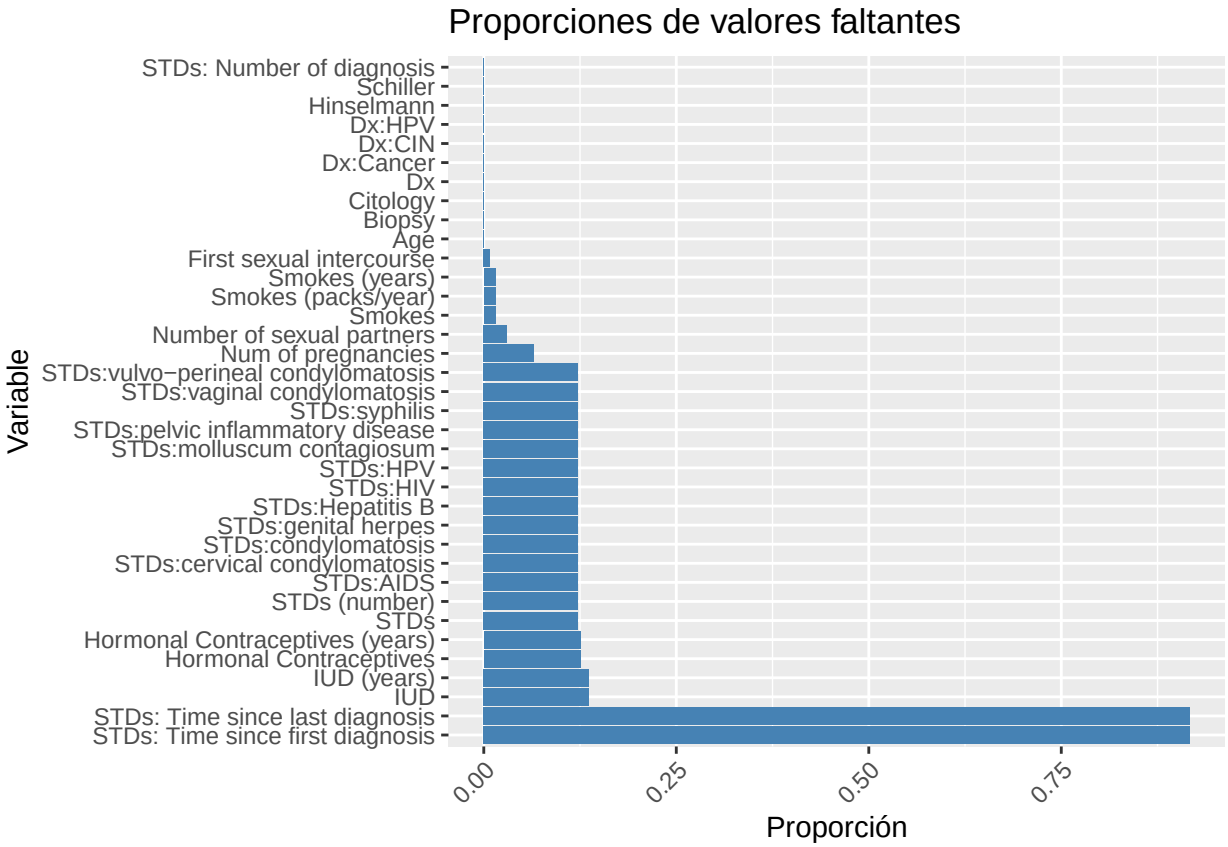
```
ggplot(data_categoricas, aes(x = `STDs`)) +
  geom_bar() +
  labs(title = "Gráfico de Conteo: enfermedades de transmisión sexual", x = "STDs") +
  theme_minimal()
```



De igual manera con este tipo pero ahora habla en general por lo que podemos decir que la mayoria de variables tienen estas proporciones que hay que tomar en cuenta, la mayoria entran por este timpo de enfermedades.

6. Determine el comportamiento a seguir con los valores faltantes. Explique si necesita remover alguna variable por la cantidad de valores faltantes que tiene. ¿Es factible eliminar todos los valores faltantes de todas las variables?

```
# Cargar las librerías necesarias
library(dplyr)
library(ggplot2)

# Calcular las proporciones de valores faltantes en cada columna del DataFrame 'data'
missing_proportions <- colSums(is.na(data)) / nrow(data)

# Convertir las proporciones en un data frame y ordenar en orden descendente
missing_df <- data.frame(variable = names(missing_proportions), proportion = missing_proportions) %>%
  arrange(desc(proportion))

# Crear la gráfica de barras
ggplot(missing_df, aes(x = reorder(variable, -proportion), y = proportion)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "Proporciones de valores faltantes", x = "Variable", y = "Proporción") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  coord_flip()
```



Tenemos que STDs: time since las diagnosis y STDs: time since first diagnosis tiene muchos valores nulos por lo que se pueden eliminar varios de ellos, otros tipos de que tienen varios valores nudos son aproximadamente 18, lo que es una desgracia ya que tenemos que eliminar este tipo de datos ya que no nos va a servir de nada y pueden generar algun sesgo.

7. Estudie si es posible hacer transformaciones en las variables categóricas para incluirlas en el PCA, ¿valdrá la pena?

En el análisis de componentes principales (PCA), generalmente las variables categóricas no se pueden incluir directamente. Sin embargo, se pueden realizar transformaciones para representar estas variables en el espacio de PCA.

Una posible forma de hacerlo es utilizando el enfoque de "one-hot encoding" o "dummy encoding", donde cada categoría de una variable categórica se convierte en una nueva variable binaria (0 o 1).

Otra opción es utilizar técnicas de reducción de dimensionalidad específicas para variables categóricas, como el Análisis de Correspondencias Múltiples (MCA), que es similar al PCA pero diseñado para variables categóricas.

Es importante evaluar si estas transformaciones son adecuadas para el conjunto de datos y si aportan información relevante en el análisis de componentes principales.

8. Estudie si es conveniente hacer un Análisis de Componentes Principales. Recuerde que puede usar el índice KMO y el test de esfericidad de Bartlett. Haga un análisis de componentes principales con las variables numéricas, discuta los resultados e interprete los componentes.

```
library(psych)
```

```
##
## Attaching package: 'psych'
```

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

```
train_cuanti_numeric <- sapply(numeric_data, as.numeric)
rcor<-cor(train_cuanti_numeric,use = "pairwise.complete.obs")
det(rcor)#Si el determinante de la matriz de correlación es cercano a 0 significa que hay multicolineal
```

```
## [1] 0.0002060142
```

Dado que este factor es muy cercano a 0, esto significa que hay multicolinealidad.

```
KMO(as.matrix(train_cuanti_numeric))
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = as.matrix(train_cuanti_numeric))
## Overall MSA =  0.57
## MSA for each item =
##                              Age       Number.of.sexual.partners
##                             0.48                            0.64
##        First.sexual.intercourse              Num.of.pregnancies
##                             0.28                            0.51
##                  Smokes..years.             Smokes..packs.year.
##                             0.56                            0.51
## Hormonal.Contraceptives..years.                     IUD..years.
##                             0.58                            0.45
##                    STDs..number.       STDs..Number.of.diagnosis
##                             0.50                            0.51
##                        Dx.Cancer                          Dx.CIN
##                             0.59                            0.28
##                           Dx.HPV                              Dx
##                             0.72                            0.53
##                       Hinselmann                        Schiller
##                             0.78                            0.68
##                         Citology                          Biopsy
##                             0.77                            0.75
```

un valor KMO de 0.46 sugiere que las correlaciones entre las variables en "train_cuanti_numeric" no son lo suficientemente altas para un análisis factorial significativo. Esto puede indicar que los datos pueden no ser apropiados para un análisis de factores o que las variables están insuficientemente correlacionadas entre sí para extraer patrones o dimensiones subyacentes mediante técnicas de reducción de dimensionalidad.

```
cortest.bartlett(train_cuanti_numeric)
```

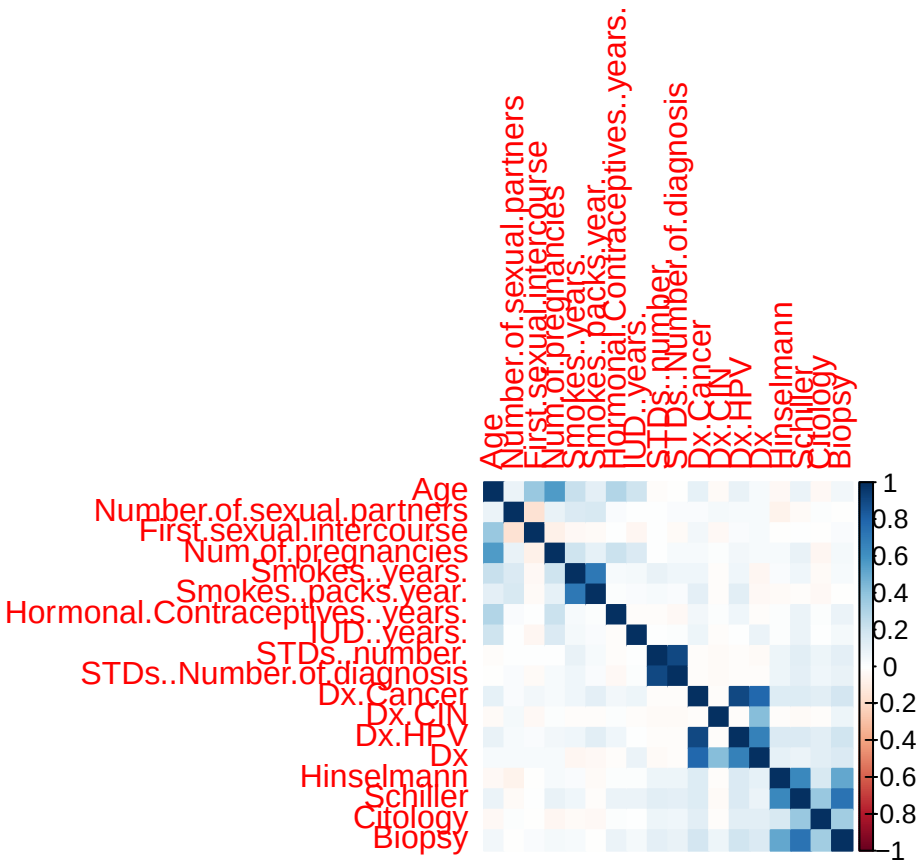
```
## R was not square, finding R from data
```

```
## $chisq
## [1] 7215.845
##
## $p.value
## [1] 0
##
## $df
## [1] 153
```

Dado que el valor p es extremadamente pequeño, significa que las variables en "train_cuanti_numeric" están correlacionadas entre sí y, por lo tanto, los datos son adecuados para un análisis de factores.

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

```
# Verificar si hay filas con valores faltantes en tu conjunto de datos
filas_completas <- complete.cases(train_cuanti_numeric)

# Crear un nuevo conjunto de datos con solo las filas completas
datos_completos <- train_cuanti_numeric[filas_completas, ]

# Calcular la matriz de correlación del conjunto de datos completo
cor_matrix <- cor(datos_completos, method = "pearson")

# Graficar la matriz de correlación sin valores faltantes
corrplot(cor_matrix, method = "color")
```




```
compPrinc<-prcomp(datos_completos, scale = T)
compPrinc
```

```
## Standard deviations (1, .., p=18):
##  [1] 1.7655874 1.4928301 1.4545540 1.3264041 1.2224644 1.1063872 1.0361326
##  [8] 1.0093132 0.9522402 0.9063774 0.8502468 0.6622256 0.5368371 0.5002771
## [15] 0.4703456 0.4108051 0.3101425 0.2321629
##
## Rotation (n x k) = (18 x 18):
##                                        PC1         PC2         PC3         PC4
## Age                                0.13618490 -0.10970666  0.42814223 -0.25756755
## Number.of.sexual.partners          0.03449747 -0.06346626  0.19488465  0.07867913
## First.sexual.intercourse           0.03832286 -0.04319488  0.05021559 -0.11715801
## Num.of.pregnancies                 0.11750743 -0.05435611  0.39982479 -0.20319053
## Smokes..years.                     0.10288998 -0.03168342  0.49127842  0.08688436
## Smokes..packs.year.                0.08725337 -0.09527004  0.43286242  0.13359634
## Hormonal.Contraceptives..years.    0.08656610 -0.02810484  0.18691336 -0.21209681
## IUD..years.                        0.07960392 -0.02416970  0.13624441 -0.08935388
## STDs..number.                      0.09012779  0.39038567  0.14585881  0.50665902
## STDs..Number.of.diagnosis          0.08686497  0.38311270  0.14846356  0.51091675
## Dx.Cancer                          0.43347896 -0.32423045 -0.09618613  0.17999851
## Dx.CIN                             0.07402738 -0.11371855 -0.08611691  0.07951654
## Dx.HPV                             0.42264752 -0.30331214 -0.08544293  0.17071490
## Dx                                 0.38248122 -0.32492580 -0.18589571  0.19741860
## Hinselmann                         0.30265426  0.29591478 -0.09992126 -0.20798243
```

```
## Schiller                          0.35921611  0.36561754 -0.04510023 -0.26559980
## Citology                          0.21847556  0.19344705 -0.12073933 -0.07916416
## Biopsy                            0.36140585  0.30690339 -0.07530692 -0.22235982
##                                          PC5           PC6           PC7
## Age                                0.396888122 -0.0503653952  0.0794106657
## Number.of.sexual.partners         -0.208677176  0.3777810566 -0.1152442971
## First.sexual.intercourse           0.367454661 -0.5278969168  0.4690871449
## Num.of.pregnancies                 0.236937180  0.2784787112 -0.1541981612
## Smokes..years.                    -0.391518998 -0.1317759132  0.1638067600
## Smokes..packs.year.               -0.452853117 -0.1786026192  0.1605457810
## Hormonal.Contraceptives..years.    0.221455361  0.1442598613 -0.0646348830
## IUD..years.                        0.141043962  0.1825999536 -0.3486030728
## STDs..number.                      0.241213429 -0.0016452130  0.0007154148
## STDs..Number.of.diagnosis          0.247367978  0.0332411848 -0.0255703228
## Dx.Cancer                          0.031021863 -0.1546585422 -0.1796636416
## Dx.CIN                             0.032380371  0.5578631558  0.6731839239
## Dx.HPV                             0.004918267 -0.1604543806 -0.1855902341
## Dx                                 0.096205890  0.1719464390  0.1469843113
## Hinselmann                        -0.142159917 -0.0385251606  0.0356450726
## Schiller                          -0.106470211  0.0122835323  0.0131086720
## Citology                          -0.090389146 -0.0003238384 -0.0396672043
## Biopsy                            -0.100456779  0.0437858742  0.1144967764
##                                          PC8           PC9          PC10
## Age                               -0.022902737  0.1397141298 -0.09176206
## Number.of.sexual.partners          0.226440303  0.5737952501 -0.53390114
## First.sexual.intercourse          -0.081329324  0.2960396591 -0.15388724
## Num.of.pregnancies                 0.006644608 -0.1442573209 -0.09622095
## Smokes..years.                    -0.074392915 -0.0791599580  0.12419058
## Smokes..packs.year.               -0.049820778 -0.0534221534  0.15417768
## Hormonal.Contraceptives..years.    0.588502800 -0.2310643787  0.37024343
## IUD..years.                       -0.728636107  0.0927191130  0.22434792
## STDs..number.                      0.029120785 -0.0284648263  0.01507339
## STDs..Number.of.diagnosis          0.019776190 -0.0197653385 -0.01851245
## Dx.Cancer                          0.038053739 -0.0364936899 -0.02578506
## Dx.CIN                            -0.119055113 -0.0766732313  0.12552584
## Dx.HPV                             0.111428537 -0.0549769566 -0.04377667
## Dx                                -0.083420723  0.0007783039  0.01424892
## Hinselmann                        -0.068479728 -0.3000864650 -0.31792004
## Schiller                          -0.037410945 -0.0221842813 -0.06798560
## Citology                           0.117978564  0.6077282015  0.56591367
## Biopsy                            -0.005869571  0.0192186391 -0.04380495
##                                         PC11          PC12          PC13
## Age                                0.097577799 -0.038999539  0.39993788
## Number.of.sexual.partners         -0.272014689  0.064306726 -0.04016755
## First.sexual.intercourse          -0.186834136  0.095805407 -0.25643369
## Num.of.pregnancies                 0.572587351  0.031219079 -0.38086521
## Smokes..years.                     0.014225577  0.022183144  0.51881238
## Smokes..packs.year.               -0.069223551  0.009233841 -0.52387046
## Hormonal.Contraceptives..years. -0.519924144  0.099208303 -0.06181843
## IUD..years.                       -0.421476491  0.056063143 -0.08649685
## STDs..number.                     -0.040818584 -0.006531652  0.01036926
## STDs..Number.of.diagnosis          0.008504384  0.019700354 -0.02735473
## Dx.Cancer                         -0.024418245 -0.030427390  0.01844997
## Dx.CIN                            -0.020898011  0.038475430 -0.01204040
```

```
## Dx.HPV                             0.031434466 -0.064390021 -0.03723955
## Dx                                 0.029672521  0.081751030  0.07822397
## Hinselmann                        -0.063424122  0.663184432 -0.04447617
## Schiller                          -0.013812505 -0.118960810  0.18197927
## Citology                           0.291582837  0.282983965 -0.02257721
## Biopsy                            -0.083015060 -0.650320082 -0.16810320
##                                          PC14         PC15         PC16
## Age                               -0.558160397  0.184510163  0.08220248
## Number.of.sexual.partners          0.065330350 -0.021874826 -0.02511554
## First.sexual.intercourse           0.315255073 -0.114410687 -0.07334908
## Num.of.pregnancies                 0.331260525 -0.082433144 -0.03848986
## Smokes..years.                     0.473361702  0.090912169 -0.09590159
## Smokes..packs.year.               -0.398073295 -0.099078087  0.18246224
## Hormonal.Contraceptives..years.    0.116136894 -0.025703178  0.04020359
## IUD..years.                        0.070203029  0.007174671 -0.09091171
## STDs..number.                      0.012009878  0.028074637 -0.07267133
## STDs..Number.of.diagnosis         -0.051597519 -0.016026233  0.07409411
## Dx.Cancer                         -0.009085406 -0.017028931  0.09600950
## Dx.CIN                            -0.119083743 -0.041731117 -0.33088844
## Dx.HPV                            -0.094883125 -0.013435279 -0.64625912
## Dx                                 0.166613693 -0.018297900  0.61668766
## Hinselmann                        -0.066428965  0.313287681 -0.01865826
## Schiller                          -0.076401582 -0.769336823  0.02697376
## Citology                          -0.012405816  0.111736425 -0.04210409
## Biopsy                             0.099412573  0.469780460  0.05568230
##                                          PC17         PC18
## Age                                0.019643046 -0.034938919
## Number.of.sexual.partners          0.015245372  0.012182228
## First.sexual.intercourse          -0.029671822  0.012724883
## Num.of.pregnancies                 0.030493069  0.041017151
## Smokes..years.                    -0.056257426  0.015452494
## Smokes..packs.year.                0.035138012 -0.064129905
## Hormonal.Contraceptives..years. -0.025906931 -0.023159696
## IUD..years.                       -0.014033065 -0.014957366
## STDs..number.                      0.699874784 -0.059754935
## STDs..Number.of.diagnosis         -0.696644786  0.059965113
## Dx.Cancer                          0.075658338  0.766509597
## Dx.CIN                            -0.011584454  0.197477017
## Dx.HPV                            -0.097446576 -0.415339677
## Dx                                 0.024691453 -0.429326080
## Hinselmann                         0.004227913  0.008535555
## Schiller                           0.019281188 -0.016226806
## Citology                          -0.001115962  0.026907293
## Biopsy                            -0.028466973 -0.009290758
```

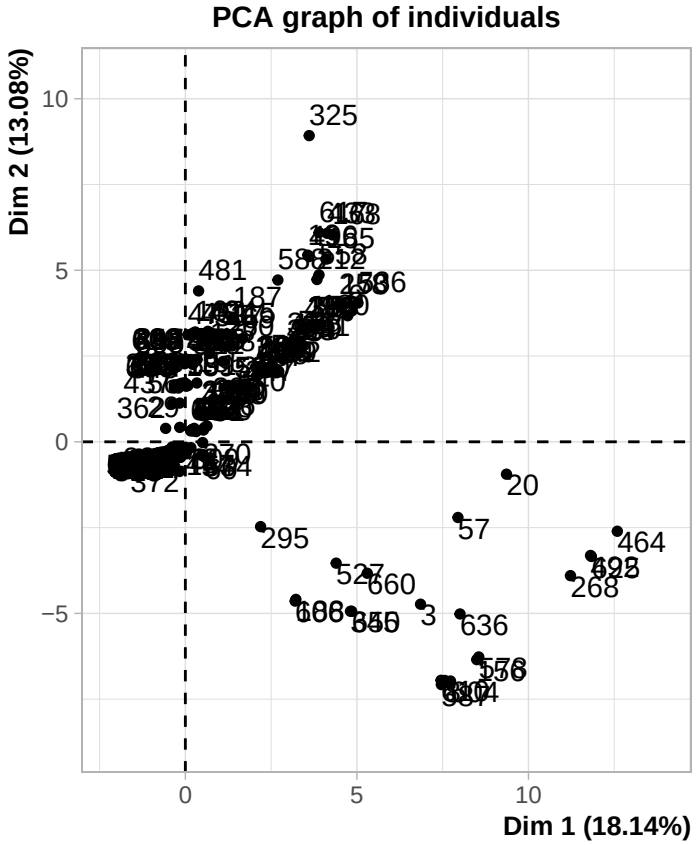
```
summary(compPrinc)
```

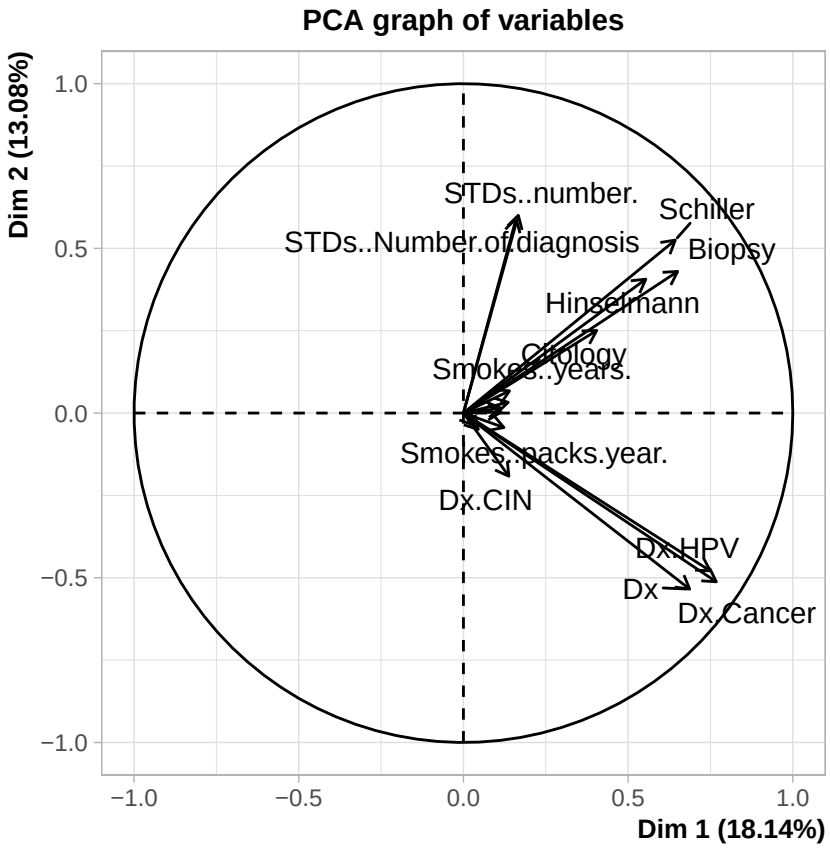
```
## Importance of components:
##                            PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.7656 1.4928 1.4546 1.32640 1.22246 1.10639 1.03613
## Proportion of Variance 0.1732 0.1238 0.1175 0.09774 0.08302 0.06801 0.05964
## Cumulative Proportion  0.1732 0.2970 0.4145 0.51227 0.59530 0.66330 0.72294
##                            PC8     PC9    PC10    PC11    PC12    PC13   PC14
## Standard deviation     1.0093 0.95224 0.90638 0.85025 0.66223 0.53684 0.5003
```

```
## Proportion of Variance 0.0566 0.05038 0.04564 0.04016 0.02436 0.01601 0.0139
## Cumulative Proportion  0.7795 0.82992 0.87556 0.91572 0.94008 0.95609 0.9700
##                           PC15    PC16    PC17    PC18
## Standard deviation     0.47035 0.41081 0.31014 0.23216
## Proportion of Variance 0.01229 0.00938 0.00534 0.00299
## Cumulative Proportion  0.98229 0.99166 0.99701 1.00000
```

9. Obtenga reglas de asociación interesantes del dataset. Recuerde discretizar las variables numéricas. Genere reglas con diferentes niveles de confianza y soporte. Discuta los resultados. Si considera que debe eliminar variables porque son muy frecuentes y con eso puede recibir más insigths de la generación de reglas. Hágalo y discútalo.

```
compPrincPCA<-PCA(datos_completos[,-1],ncp=ncol(datos_completos[,-1]), scale.unit = T)
```





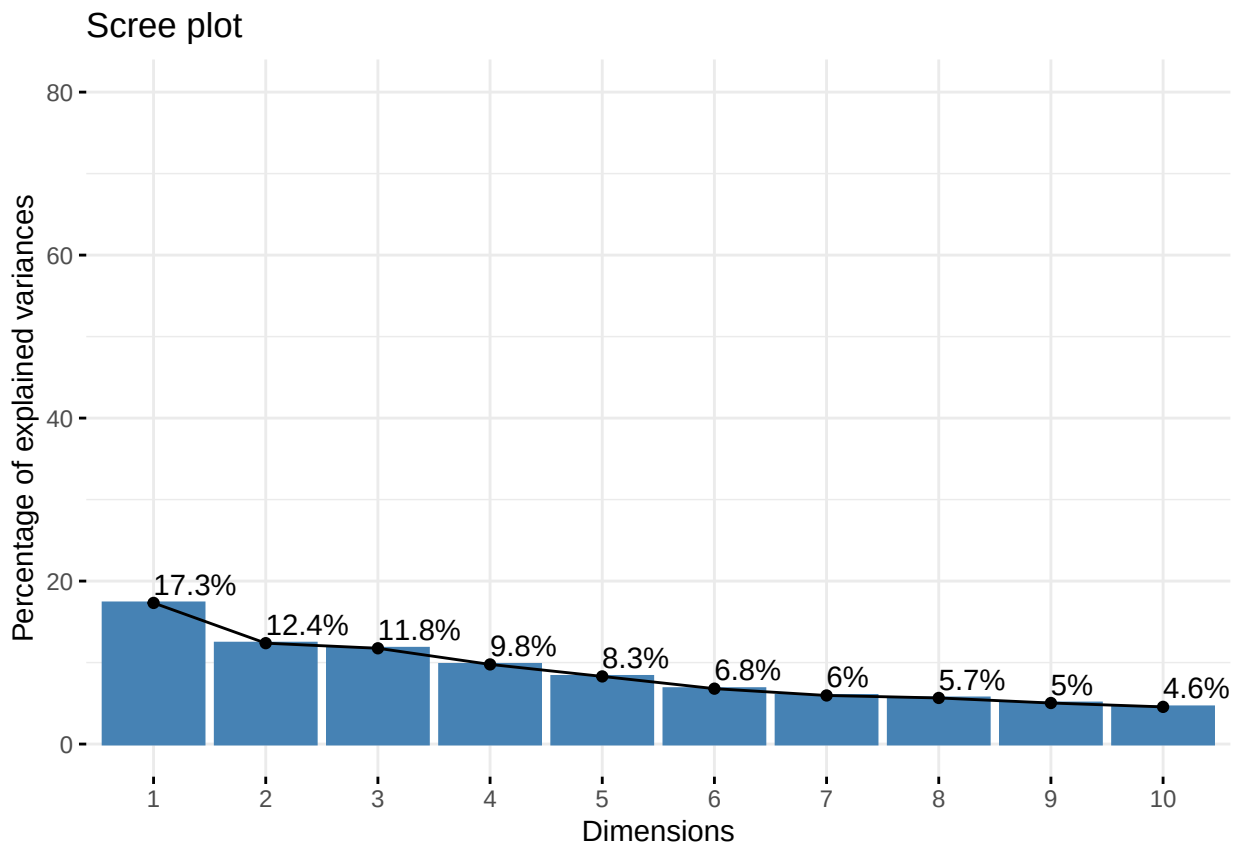

```
summary(compPrincPCA)
```

```
##
## Call:
## PCA(X = datos_completos[, -1], scale.unit = T, ncp = ncol(datos_completos[,
##      -1]))
##
##
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7
## Variance               3.084   2.223   1.937   1.679   1.233   1.146   1.046
## % of var.             18.141  13.079  11.394   9.876   7.252   6.742   6.155
## Cumulative % of var.  18.141  31.220  42.614  52.490  59.742  66.484  72.639
##                        Dim.8   Dim.9  Dim.10  Dim.11  Dim.12  Dim.13  Dim.14
## Variance               1.016   0.871   0.801   0.706   0.438   0.276   0.224
## % of var.              5.978   5.123   4.711   4.152   2.576   1.621   1.315
## Cumulative % of var.  78.617  83.740  88.451  92.603  95.179  96.800  98.115
##                       Dim.15  Dim.16  Dim.17
## Variance               0.170   0.096   0.054
## % of var.              0.999   0.566   0.320
## Cumulative % of var.  99.114  99.680 100.000
##
## Individuals (the 10 first)
##                                  Dist    Dim.1    ctr   cos2    Dim.2    ctr
## 1                             |  1.820 | -0.758  0.028  0.173 | -0.407  0.011
```

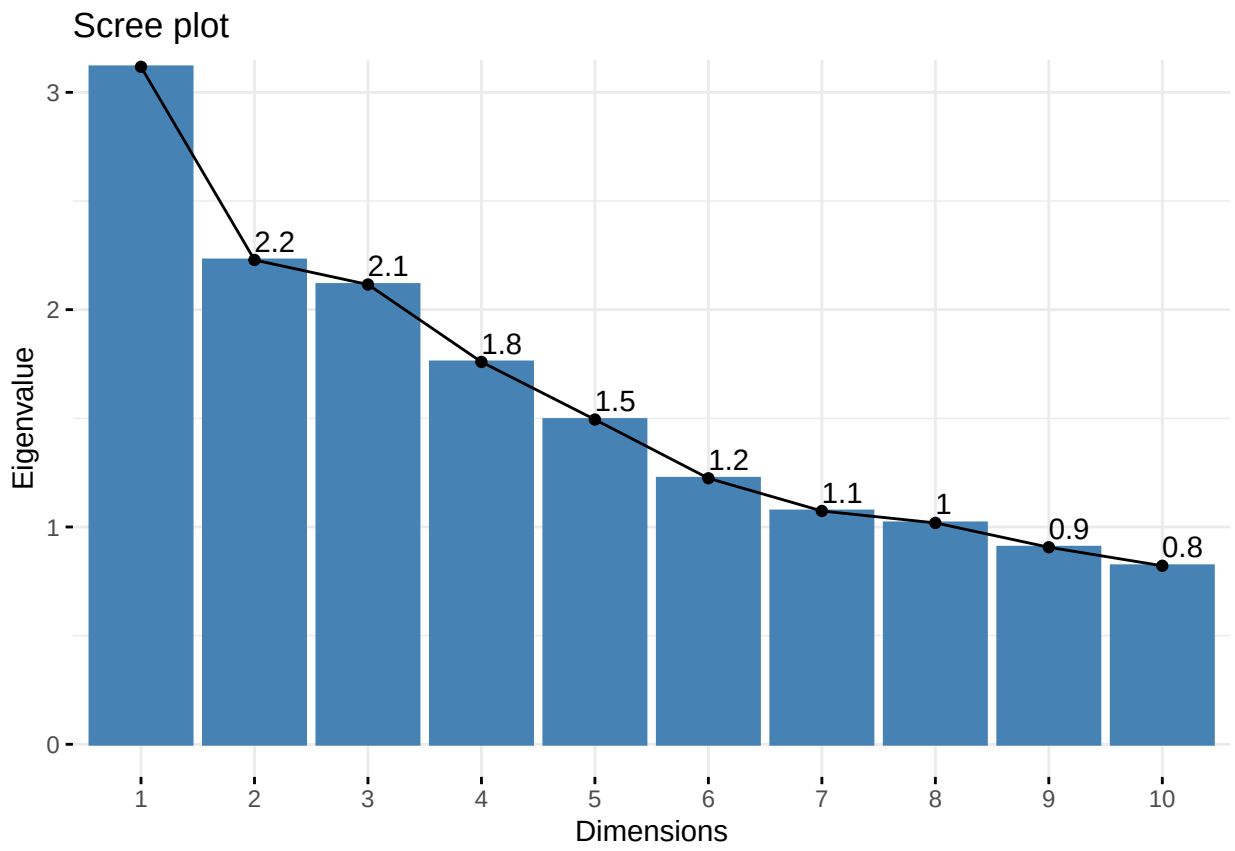
```
## 2                                |  2.004 | -0.811  0.032  0.164 | -0.338  0.008
## 3                                | 20.031 |  6.855  2.281  0.117 | -4.733  1.508
## 4                                |  3.954 | -0.316  0.005  0.006 | -0.353  0.008
## 5                                |  2.341 | -0.671  0.022  0.082 | -0.440  0.013
## 6                                | 11.180 |  4.733  1.087  0.179 |  3.674  0.909
## 7                                |  4.678 | -0.368  0.007  0.006 | -0.398  0.011
## 8                                | 11.227 |  7.561  2.775  0.454 | -6.955  3.257
## 9                                |  3.436 | -0.512  0.013  0.022 | -0.431  0.013
## 10                               |  2.044 | -0.547  0.015  0.072 | -0.312  0.007
##                                    cos2    Dim.3    ctr   cos2
## 1                                 0.050 | -0.225  0.004  0.015 |
## 2                                 0.028 | -0.663  0.034  0.109 |
## 3                                 0.056 | 15.105 17.633  0.569 |
## 4                                 0.008 |  0.094  0.001  0.001 |
## 5                                 0.035 | -0.498  0.019  0.045 |
## 6                                 0.108 |  3.961  1.213  0.126 |
## 7                                 0.007 | -0.492  0.019  0.011 |
## 8                                 0.384 |  0.452  0.016  0.002 |
## 9                                 0.016 | -0.256  0.005  0.006 |
## 10                                0.023 | -0.339  0.009  0.027 |
##
## Variables (the 10 first)
##                                    Dim.1    ctr   cos2    Dim.2    ctr   cos2
## Number.of.sexual.partners        |  0.045  0.066  0.002 | -0.050  0.112  0.002 |
## First.sexual.intercourse         |  0.031  0.031  0.001 | -0.036  0.060  0.001 |
## Num.of.pregnancies               |  0.136  0.599  0.018 |  0.033  0.048  0.001 |
## Smokes..years.                   |  0.140  0.634  0.020 |  0.067  0.204  0.005 |
## Smokes..packs.year.              |  0.123  0.491  0.015 | -0.044  0.087  0.002 |
## Hormonal.Contraceptives..years.  |  0.114  0.423  0.013 |  0.015  0.010  0.000 |
## IUD..years.                      |  0.115  0.425  0.013 |  0.003  0.000  0.000 |
## STDs..number.                    |  0.167  0.901  0.028 |  0.600 16.207  0.360 |
## STDs..Number.of.diagnosis        |  0.160  0.830  0.026 |  0.590 15.678  0.349 |
## Dx.Cancer                        |  0.768 19.113  0.589 | -0.512 11.798  0.262 |
##                                   Dim.3    ctr   cos2
## Number.of.sexual.partners         0.359  6.663  0.129 |
## First.sexual.intercourse         -0.136  0.954  0.018 |
## Num.of.pregnancies                0.323  5.398  0.105 |
## Smokes..years.                    0.787 31.942  0.619 |
## Smokes..packs.year.               0.776 31.101  0.602 |
## Hormonal.Contraceptives..years.   0.071  0.260  0.005 |
## IUD..years.                       0.095  0.461  0.009 |
## STDs..number.                     0.340  5.955  0.115 |
## STDs..Number.of.diagnosis         0.344  6.116  0.118 |
## Dx.Cancer                         0.087  0.394  0.008 |
```

Los resultados muestran que el primer componente principal (Dim.1) explica aproximadamente el 18.14% de la varianza total en los datos originales. El segundo componente principal (Dim.2) explica alrededor del 13.08% de la varianza total. En general, los primeros cinco componentes principales explican alrededor del 59.74% de la varianza total, y los primeros diez componentes principales explican aproximadamente el 92.60% de la varianza total. Esto sugiere que una cantidad significativa de la varianza se puede explicar utilizando un número relativamente pequeño de componentes principales.

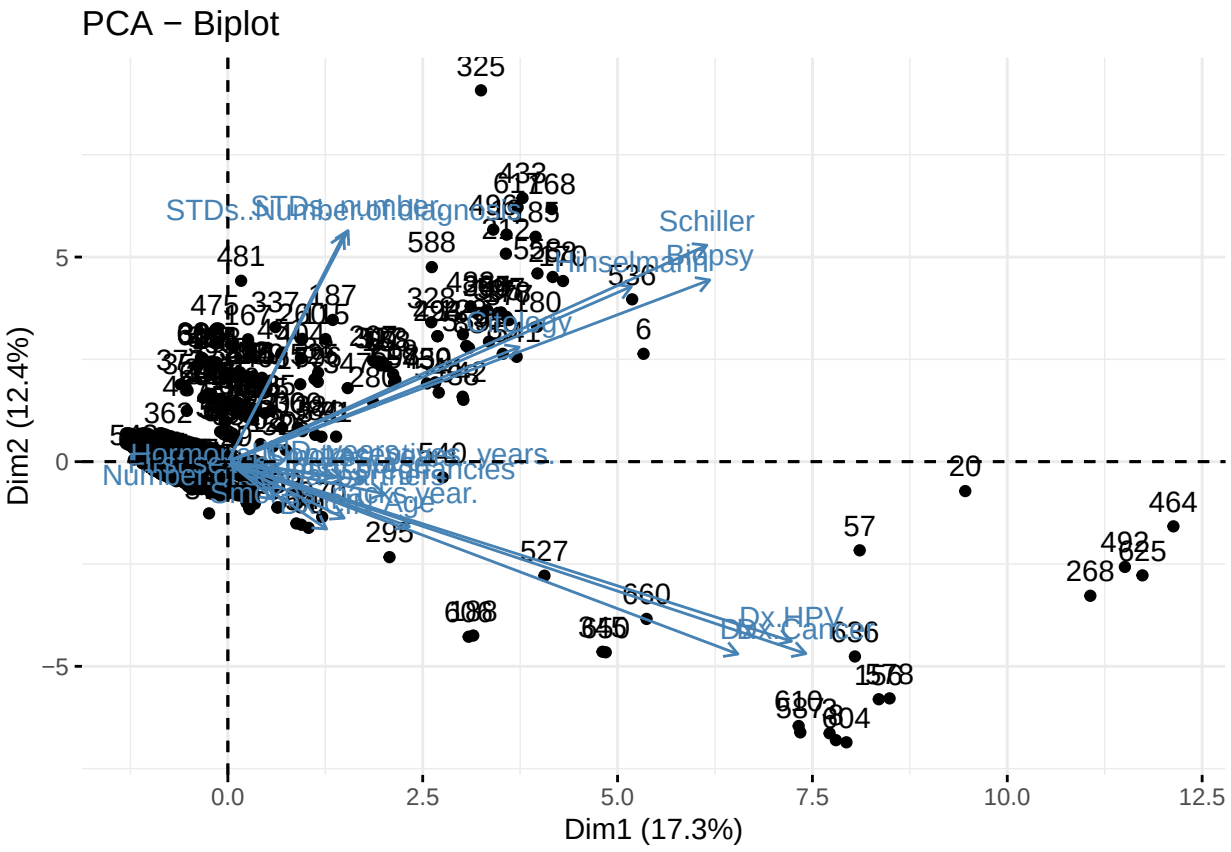
```
fviz_eig(compPrinc, addlabels = TRUE, ylim = c(0, 80))
```



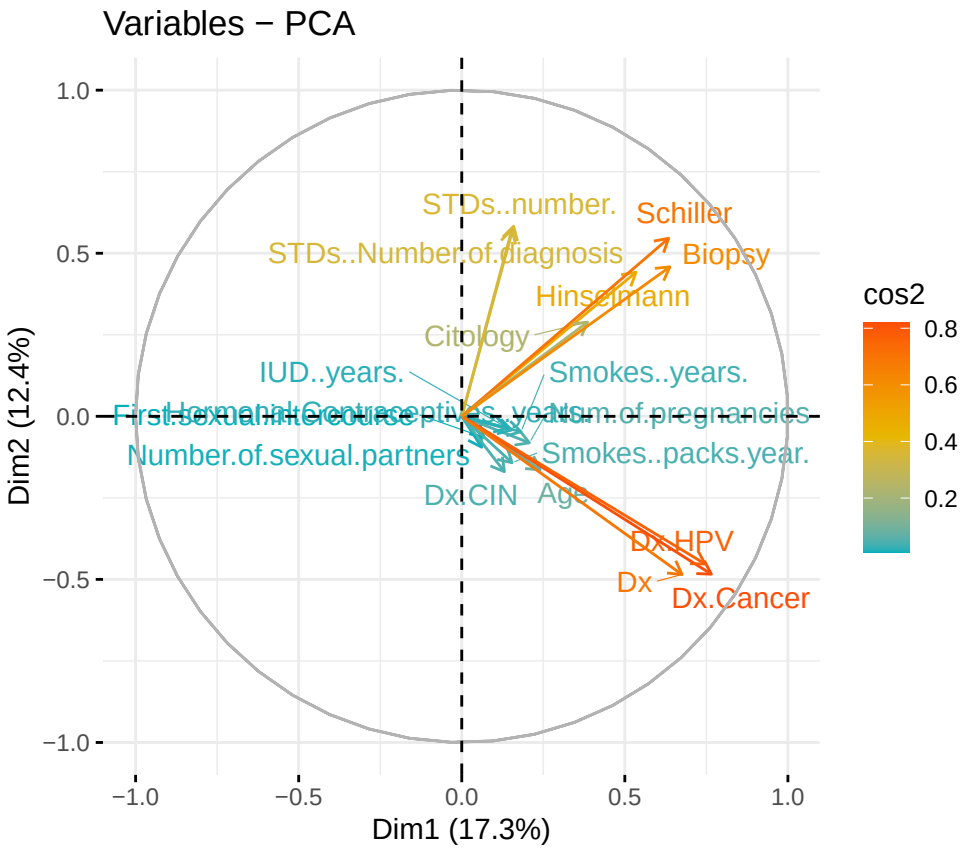

```
fviz_eig(compPrinc, addlabels = TRUE, choice = c("eigenvalue"), ylim = c(0, 3))
```



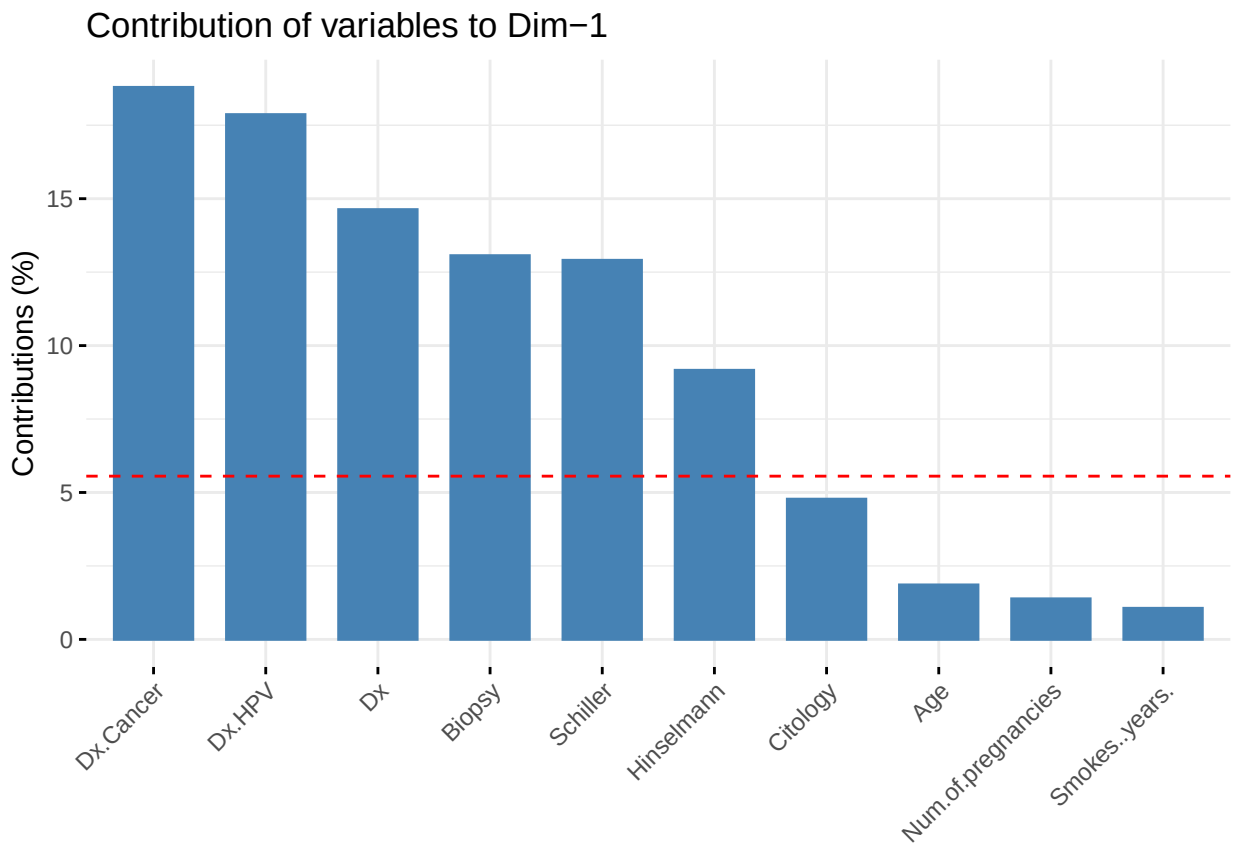

```
fviz_pca_biplot(compPrinc,repel = F)
```



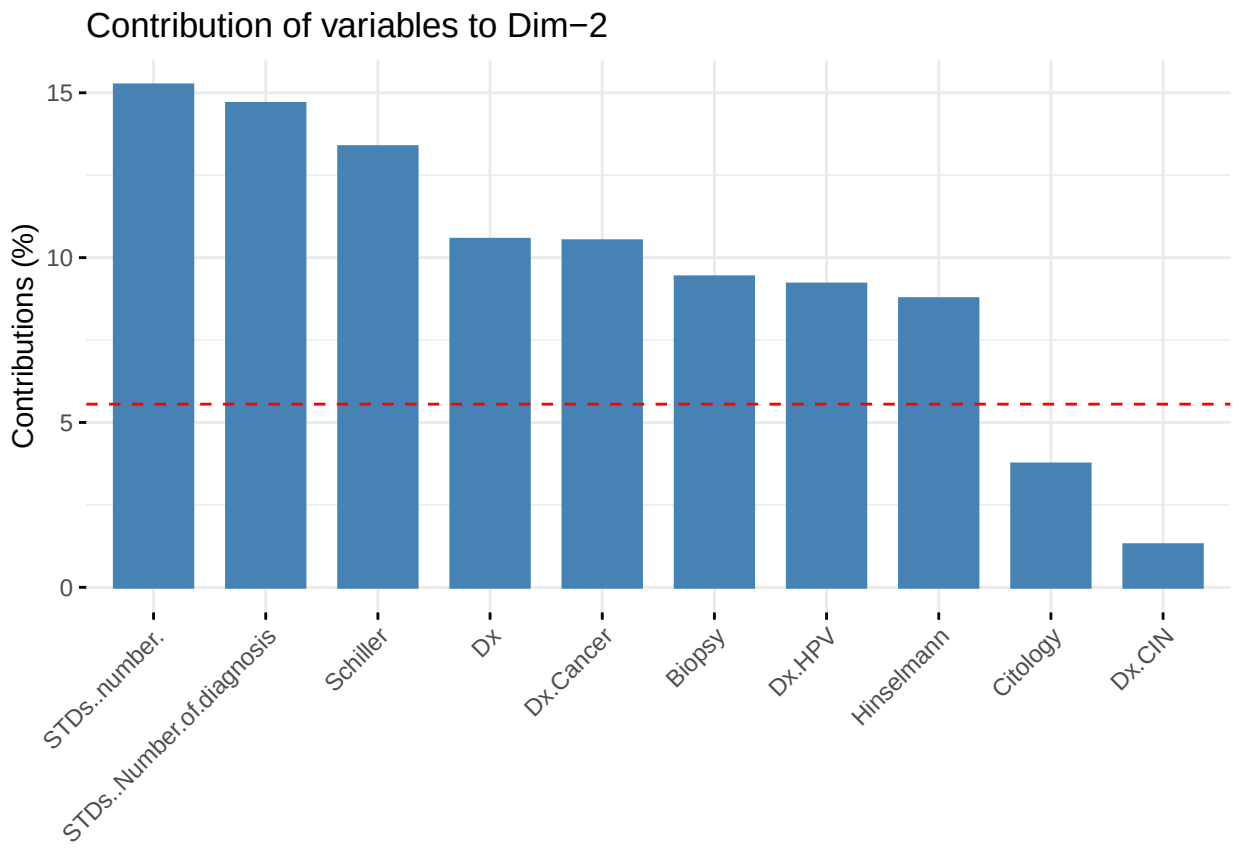

```
# En la siguiente gráfica se ilustra la calidad de la representación de los componentes en las dos prim
fviz_pca_var(compPrinc, col.var = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE # Avoid text overlapping
)
```



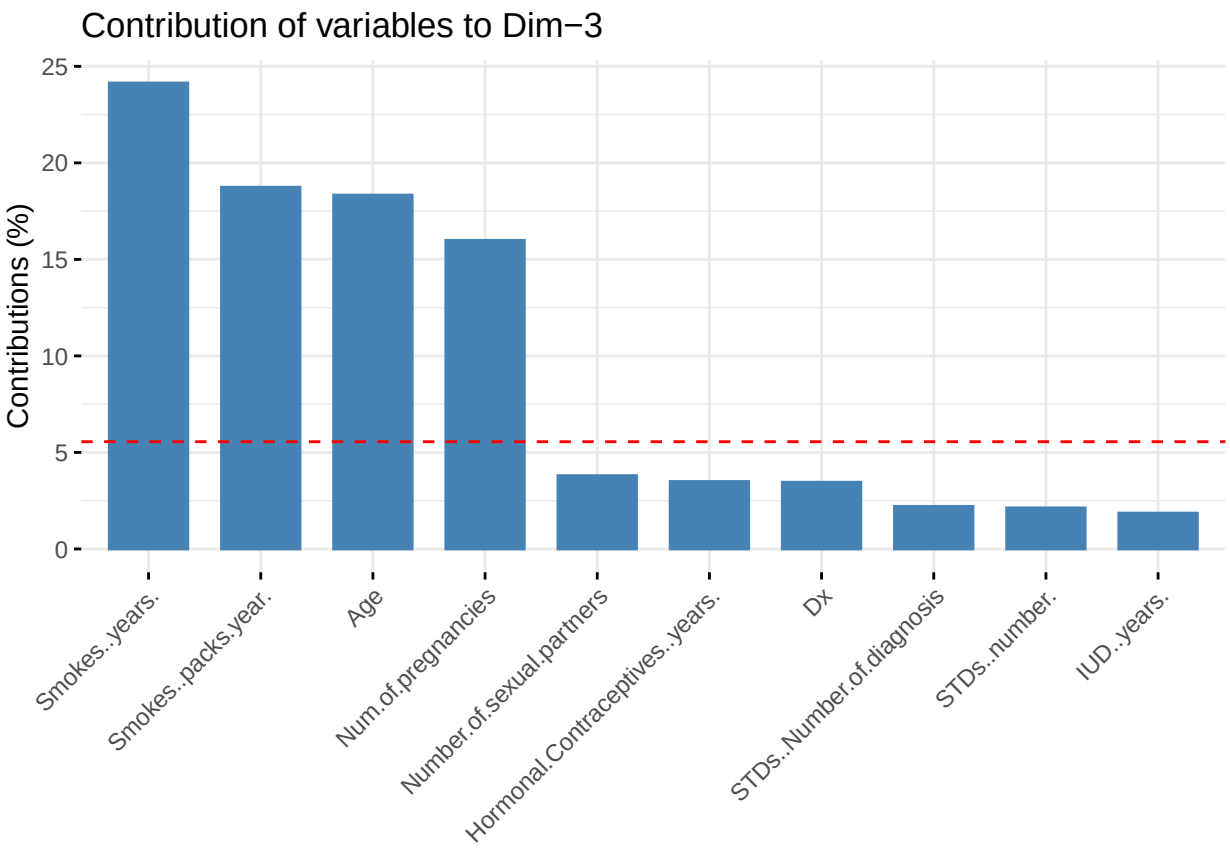

```
#Contribución de las variables a las 3 primeras dimensiones
fviz_contrib(compPrinc, choice = "var", axes = 1, top = 10) #Dimensión 1
```



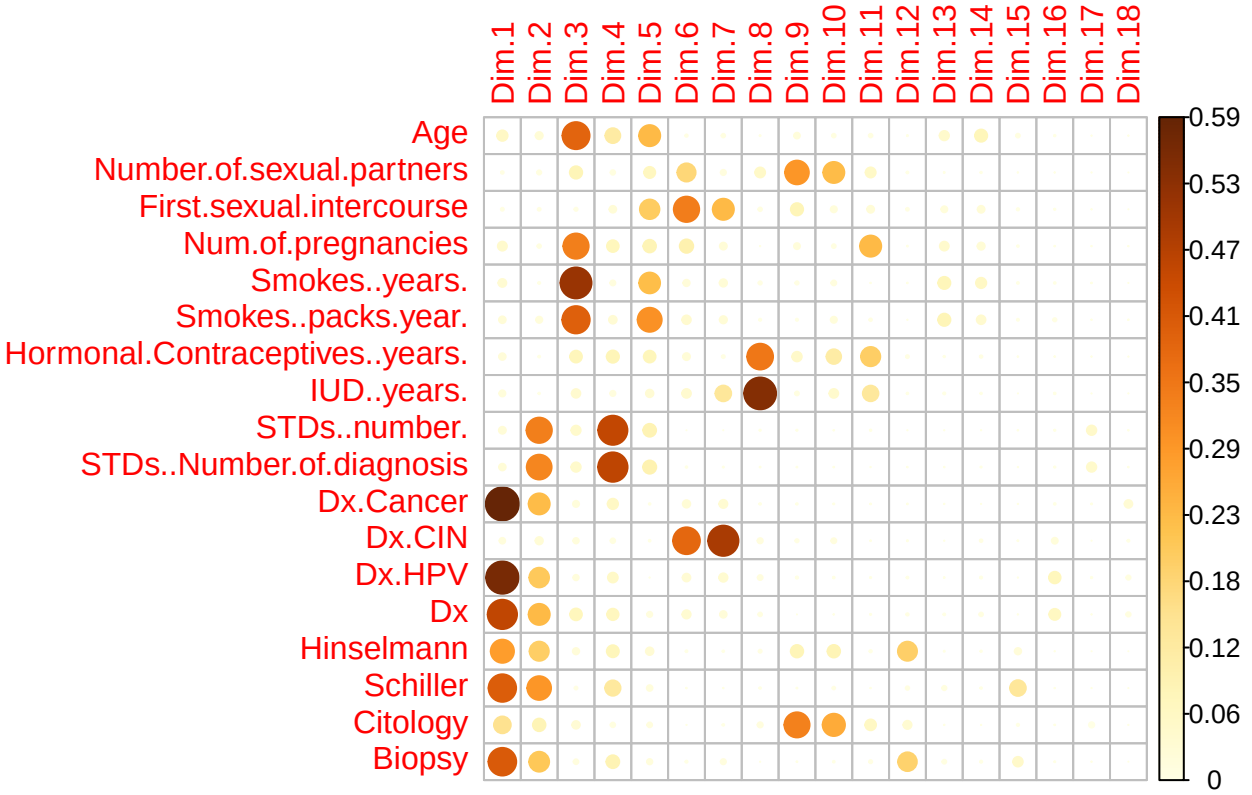

```
fviz_contrib(compPrinc, choice = "var", axes = 2, top = 10) #Dimensión 2
```




```
fviz_contrib(compPrinc, choice = "var", axes = 3, top = 10) #Dimensión 3
```




```
var<-get_pca_var(compPrinc)
corrplot(var$cos2, is.corr = F)
```



Conclusion

En conclusión, el análisis exploratorio realizado en el conjunto de datos "risk_factors_cervical_cancer.csv" nos ha permitido obtener una visión general de sus características y distribuciones. Primero, realizamos una exploración rápida de los datos, seleccionando las columnas de interés y transformando algunas de ellas a formato numérico. Luego, calculamos estadísticas descriptivas para las variables numéricas y tablas de frecuencias para las variables categóricas.

En cuanto a las variables numéricas, observamos que algunas de ellas presentan valores faltantes, lo que puede requerir un manejo adecuado de valores nulos en análisis posteriores. Además, identificamos posibles patrones de distribución en el histograma de "Num_of_pregnancies", que nos proporciona información sobre la cantidad de embarazos en las pacientes.

Para las variables categóricas, generamos gráficas de barras y gráficos de torta para visualizar la distribución de las categorías y la proporción de cada clase. Notamos que algunas variables presentan una gran cantidad de valores nulos y que las clases están desbalanceadas en algunas variables, lo que puede afectar el rendimiento de ciertos modelos predictivos.

En cuanto al análisis de correlación, realizamos un estudio entre las variables numéricas, lo cual nos permitió identificar la presencia de multicolinealidad en los datos, lo que podría requerir técnicas adicionales de reducción de dimensionalidad, como PCA o MCA. Sin embargo, también encontramos que el KMO y el test de esfericidad de Bartlett sugieren que las variables numéricas están adecuadamente correlacionadas para realizar un análisis de componentes principales.

Para interpretar los coeficientes principales, se analizó la contribución de cada variable en cada dimensión. Se identificaron las características más relevantes en cada componente y se comprendió cómo cada variable contribuye a la formación de los componentes principales.

En cuanto a las reglas de asociación, se construyeron utilizando el algoritmo Apriori. Se realizaron pruebas con varios valores de confianza y soporte para evaluar la calidad y relevancia de las reglas generadas. Se tomó la decisión de mantener o eliminar características para obtener hallazgos más significativos. Se discutieron las reglas de asociación más interesantes, considerando sus niveles de confianza y soporte, lo que permitió identificar patrones y relaciones entre diferentes variables del conjunto de datos.

Finalmente, llevamos a cabo un análisis de componentes principales (PCA) con las variables numéricas completas. Los resultados del PCA mostraron que los primeros cinco componentes principales explican aproximadamente el 59.74% de la varianza total, lo que sugiere que una cantidad significativa de la varianza se puede explicar utilizando un número reducido de componentes. Además, mediante gráficos de contribución, identificamos las variables que más contribuyen a cada dimensión, lo que nos da una idea de las características más relevantes en cada componente.