Analisis_Descriptivo_Dropout

2023-03-04

¿Los alumnos dejan la carrera?

```
datos <- read.csv("dropout.csv")
library(ggplot2)
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
```

```
#library(plotly)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

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

```
library(caret)
```

```
## Loading required package: lattice
```

```
library(rpart.plot)
```

```
## Loading required package: rpart
```

```
library(class)
library(gtable)
#library(ggmosaic)
library(ggridges)
library(fmsb)
```

```
## Registered S3 methods overwritten by 'fmsb':
##   method    from
##   print.roc pROC
##   plot.roc  pROC
```

```
set.seed(125)
```

Organizo el dataset

```
datos$Target <- as.factor(datos$Target)

datos <- datos %>%
  mutate(Gender = case_when(
    Gender == 0 ~ "mujer",
    Gender == 1 ~ "hombre",
    TRUE ~ NA_character_
  ))
datos$Gender <- as.factor(datos$Gender)


datos <- datos %>%
  mutate(Marital.status = case_when(
    Marital.status == 1 ~ "single",
    Marital.status == 2 ~ "married",
    Marital.status == 3 ~ "widower",
    Marital.status == 4 ~ "divorced",
    Marital.status == 5 ~ "facto union",
    Marital.status == 6 ~ "legally separated",
    TRUE ~ NA_character_
  ))

datos$Marital.status <- as.factor(datos$Marital.status)

datos <- datos %>%
  mutate(Course = case_when(
    Course == 1 ~ "Biofuel Production Technologies",
    Course == 2 ~ "Animation and Multimedia Design",
    Course == 3 ~ "Social Service (evening attendance)",
    Course == 4 ~ "Agronomy",
    Course == 5 ~ "Communication Design",
    Course == 6 ~ "Veterinary Nursing",
    Course == 7 ~ "Informatics Engineering",
    Course == 8 ~ "Equinculture",
    Course == 9 ~ "Management",
    Course == 10 ~ "Social Service",
    Course == 11 ~ "Tourism",
    Course == 12 ~ "Nursing",
    Course == 13 ~ "Oral Hygiene",
    Course == 14 ~ "Advertising and Marketing Management",
    Course == 15 ~ "Journalism and Communication",
    Course == 16 ~ "Basic Education",
    Course == 17 ~ "Management",
    TRUE ~ NA_character_
  ))

datos$Course <- as.factor(datos$Course)


datos <- datos %>%
  mutate(Nacionality = case_when(
    Nacionality == 1 ~ "Portuguese",
    Nacionality == 2 ~ "German",
    Nacionality == 3 ~ "Spanish",
    Nacionality == 4 ~ "Italian",
    Nacionality == 5 ~ "Dutch",
    Nacionality == 6 ~ "English",
    Nacionality == 7 ~ "Lithuanian",
    Nacionality == 8 ~ "Angolan",
    Nacionality == 9 ~ "Cape Verdean",
    Nacionality == 10 ~ "Guinean",
    Nacionality == 11 ~ "Mozambican",
    Nacionality == 12 ~ "Santomean",
    Nacionality == 13 ~ "Turkish",
    Nacionality == 14 ~ "Brazilian",
    Nacionality == 15 ~ "Romanian",
    Nacionality == 16 ~ "Moldova",
    Nacionality == 17 ~ "Mexican",
    Nacionality == 18 ~ "Ukrainian",
    Nacionality == 19 ~ "Russian",
    Nacionality == 20 ~ "Cuban",
    Nacionality == 21 ~ "Colombian",
    TRUE ~ NA_character_
  ))

datos$Displaced <- as.logical(datos$Displaced)
datos$Educational.special.needs <- as.logical(datos$Educational.special.needs)
datos$Debtor <- as.logical(datos$Debtor)
datos$Tuition.fees.up.to.date <- as.logical(datos$Tuition.fees.up.to.date)
datos$Scholarship.holder <- as.logical(datos$Scholarship.holder)
datos$International <- as.logical(datos$International)
datos$Daytime.attendance <- as.logical(datos$Daytime.evening.attendance)


bool_cols <- c("Displaced", "Educational.special.needs", "Debtor", "Tuition.fees.up.to.date", "Scholars
datos[bool_cols] <- lapply(datos[bool_cols], factor)

datos$Nacionality <- as.factor(datos$Nacionality)

datos$Gender_Target <- paste(datos$Gender, datos$Target, sep = " ")
datos$Gender_Target <- gsub(" ", "_", datos$Gender_Target)
```

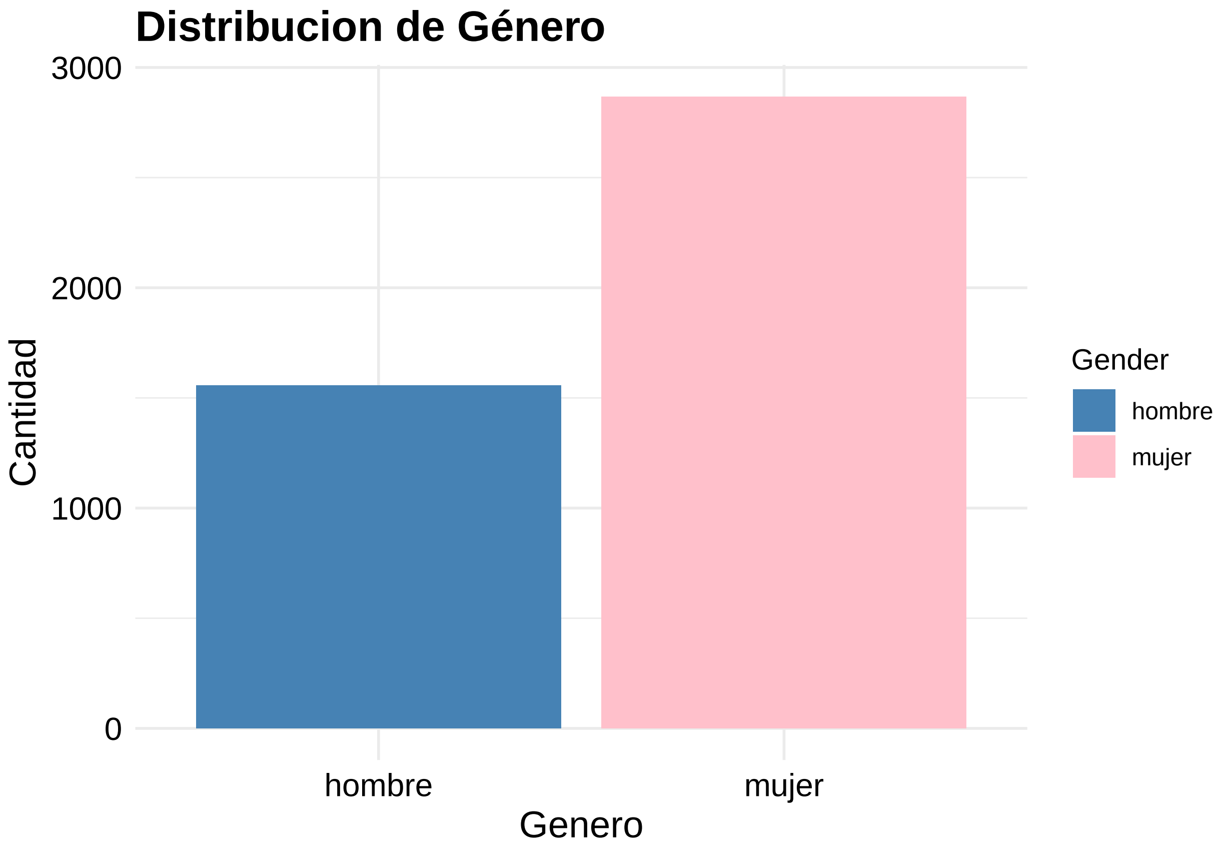
```
occupation_map_mujeres <- c("Secondary Degree",
                    "Bachelor's Degree",
                    "Bachelor's Degree",
                    "Master's",
```

```r
                    "Doctorate",
                    NA,
                    NA,
                    NA,
                    "Uncompleted secondary school",
                    "Uncompleted secondary school",
                    NA,
                    "Uncompleted secondary school",
                    "Specialized course",
                    "Uncompleted secondary school",
                    NA,
                    NA,
                    NA,
                    "Specialized course",
                    "Uncompleted secondary school",
                    NA,
                    "Specialized course",
                    NA,
                    NA,
                    NA,
                    "Uncompleted primary school",
                    "Uncompleted secondary school",
                    NA,
                    "Uncompleted secondary school",
                    NA,
                    NA,
                    "Uncompleted primary school",
                    "Uncompleted primary school",
                    "Uncompleted primary school",
                    "Uncompleted primary school",
                    "Specialized course",
                    "Secondary Degree",
                    "Specialized course",
                    "Specialized course",
                    "Bachelor's Degree",
                    "Master's")

datos$Mother.s.occupation.name <- occupation_map_mujeres[datos$Mother.s.occupation]
datos$Mother.s.occupation.name <- as.factor(datos$Mother.s.occupation.name)


occupation_map_hombres <- c("Secondary Degree",
                    "Bachelor's Degree",
                    "Bachelor's Degree",
                    "Master's",
                    "Doctorate",
                    NA,
                    NA,
                    NA,
                    "Uncompleted secondary school",
                    "Uncompleted secondary school",
                    NA,
                    "Uncompleted secondary school",
                    "Specialized course",
                    "Uncompleted secondary school",
                    NA,
                    NA,
                    NA,
                    "Specialized course",
                    "Uncompleted secondary school",
                    NA,
                    "Specialized course",
                    NA,
                    NA,
                    NA,
                    "Uncompleted primary school",
                    "Uncompleted secondary school",
                    NA,
                    "Uncompleted secondary school",
                    NA,
                    NA,
                    "Uncompleted primary school",
                    "Uncompleted primary school",
                    "Uncompleted primary school",
                    "Uncompleted primary school",
                    "Specialized course",
                    "Secondary Degree",
                    "Specialized course",
                    "Specialized course",
                    "Bachelor's Degree",
                    "Master's")

datos$Father.s.occupation.name <- occupation_map_hombres[datos$Father.s.occupation]
datos$Father.s.occupation.name <- as.factor(datos$Father.s.occupation.name)

datos$promedioNotas <- (datos$Curricular.units.1st.sem..grade. + datos$Curricular.units.2nd.sem..grade.)


datos_mujer <- datos %>% filter(Gender == "mujer")
datos_hombre <- datos %>% filter(Gender == "hombre")
datos_scholarship <- datos %>% filter(Scholarship.holder == TRUE)
datos_NOTscholarship <- datos %>% filter(Scholarship.holder == FALSE)
```

Análisis descriptivo

Por género

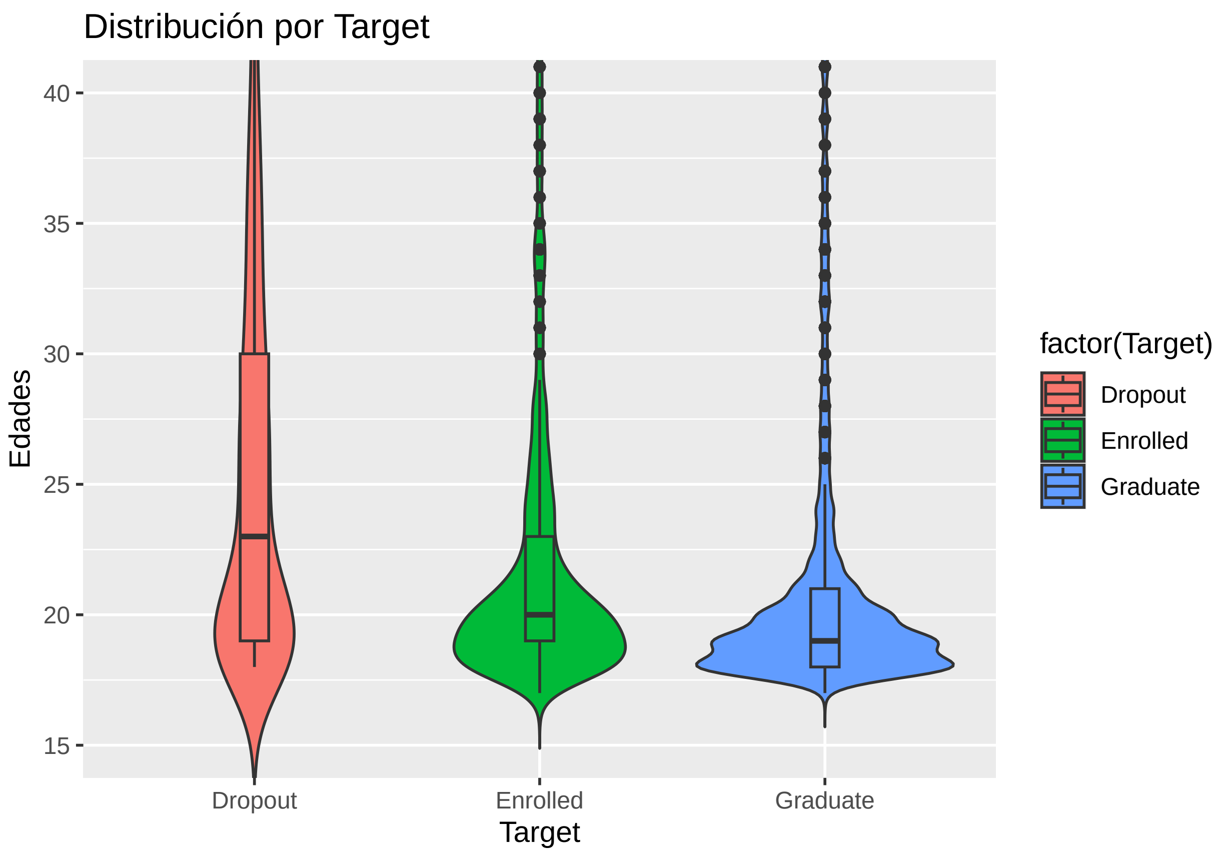
```
ggplot(datos, aes(x = Gender, fill = Gender)) +
  geom_bar() +
  labs(x = "Genero", y = "Cantidad", title = "Distribucion de Género") +
  scale_fill_manual(values = c("steelblue", "pink")) +
  theme_minimal() +
  theme(axis.text = element_text(size = 12, color = "black"),
        axis.title = element_text(size = 14),
        plot.title = element_text(size = 16, face = "bold"))
```



Por edades

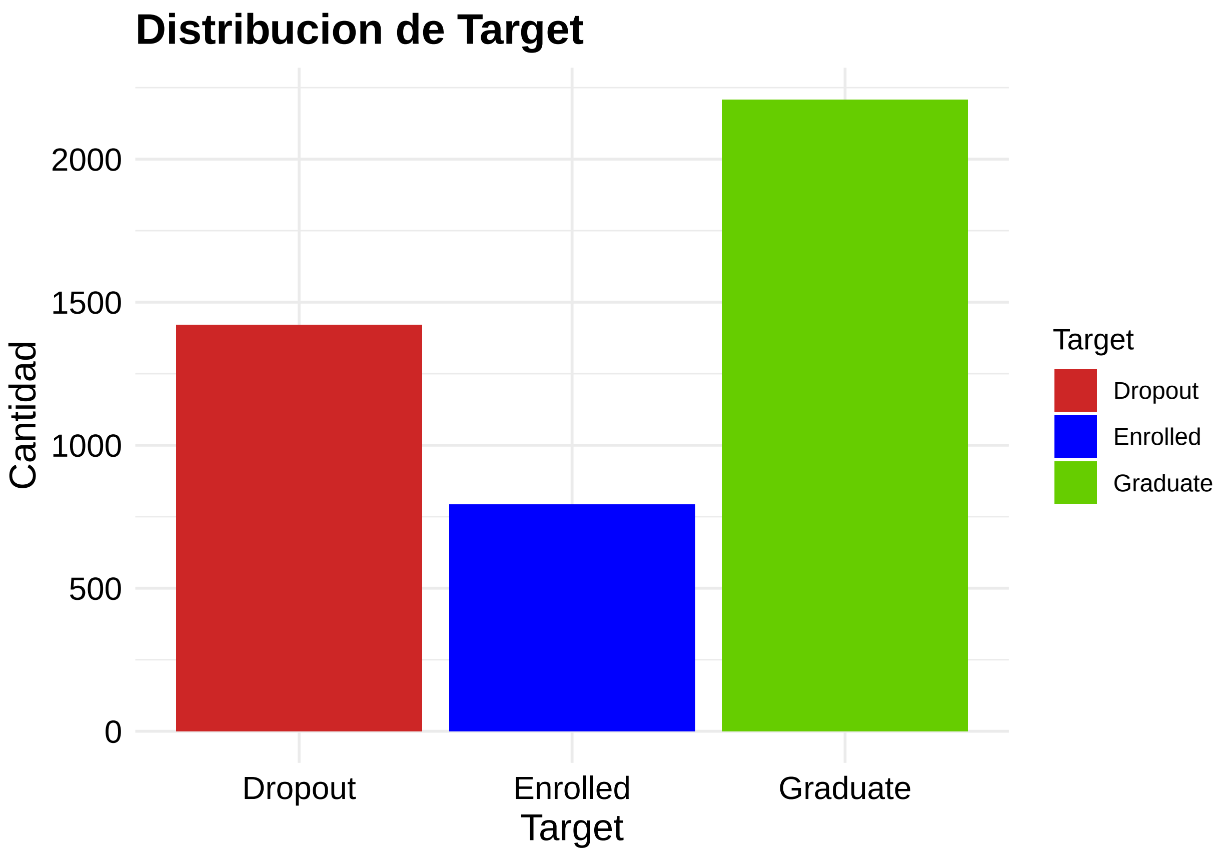
```
qplot(Target, Age.at.enrollment, data=datos,
 geom=c("violin"), trim = FALSE,  fill = factor(Target))+
  geom_boxplot(width=0.1)+
  labs(y = "Edades", x = "Target", title = "Distribución por Target")+
  coord_cartesian(ylim = c(15, 40))
```

```
## Warning: ‘qplot()‘ was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



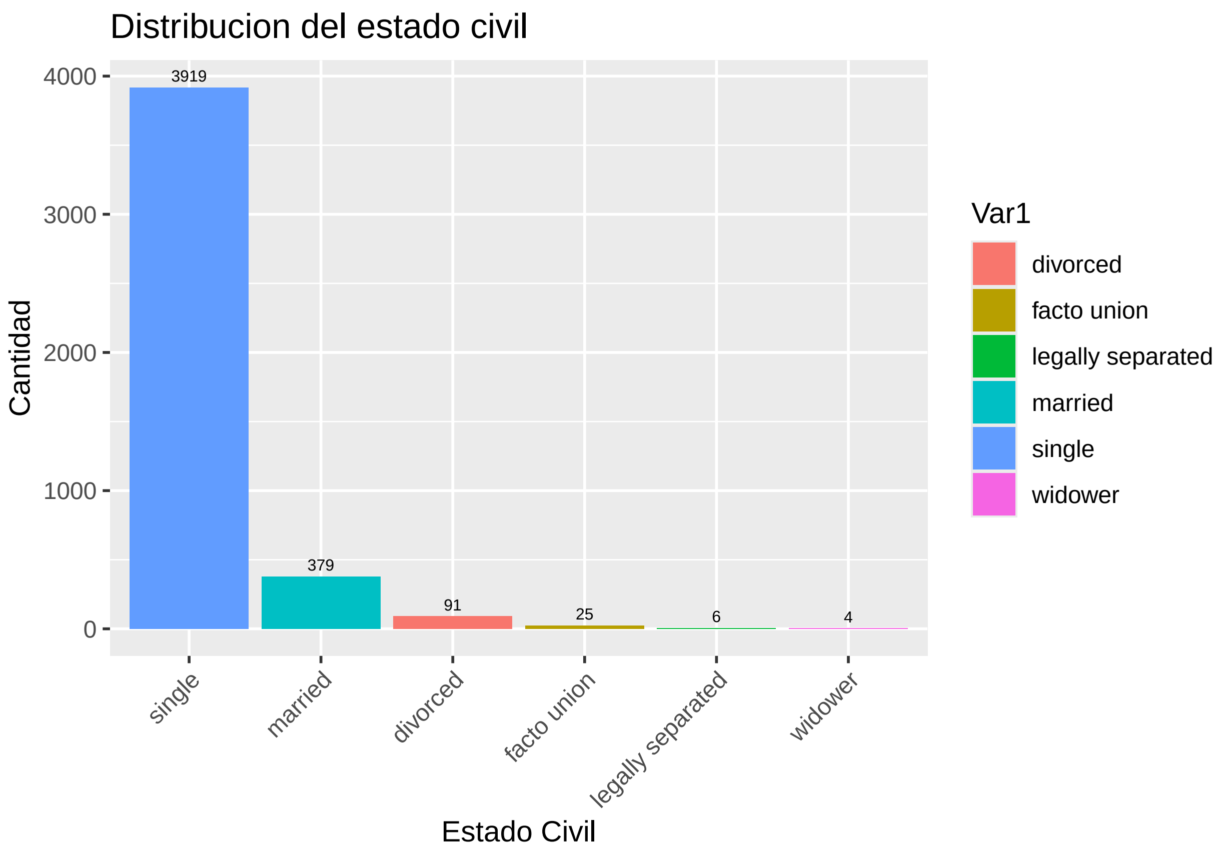
Cantidad de alumnos por target

```
ggplot(datos, aes(x = Target, fill = Target)) +
  geom_bar() +
  labs(x = "Target", y = "Cantidad", title = "Distribucion de Target") +
  scale_fill_manual(values = c("firebrick3", "blue", "chartreuse3")) +
  theme_minimal() +
  theme(axis.text = element_text(size = 12, color = "black"),
        axis.title = element_text(size = 14),
        plot.title = element_text(size = 16, face = "bold"))
```



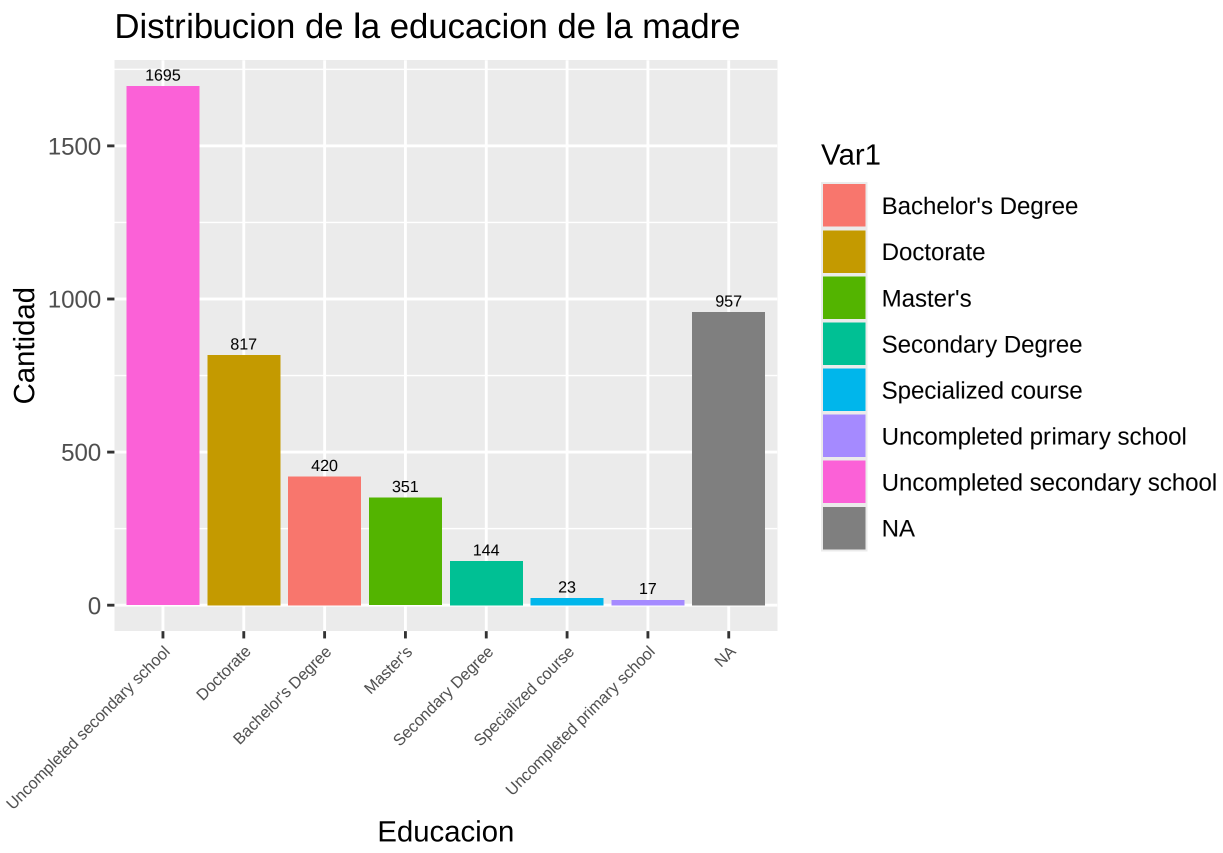
Por estado civil

```
contarMaritalStatus <- data.frame(table(datos$Marital.status))

ggplot(contarMaritalStatus, aes(x = reorder(Var1, -Freq), y = Freq, fill = Var1)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = Freq), vjust = -0.5, size = 2) +
  labs(x = "Estado Civil", y = "Cantidad", title = "Distribucion del estado civil") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



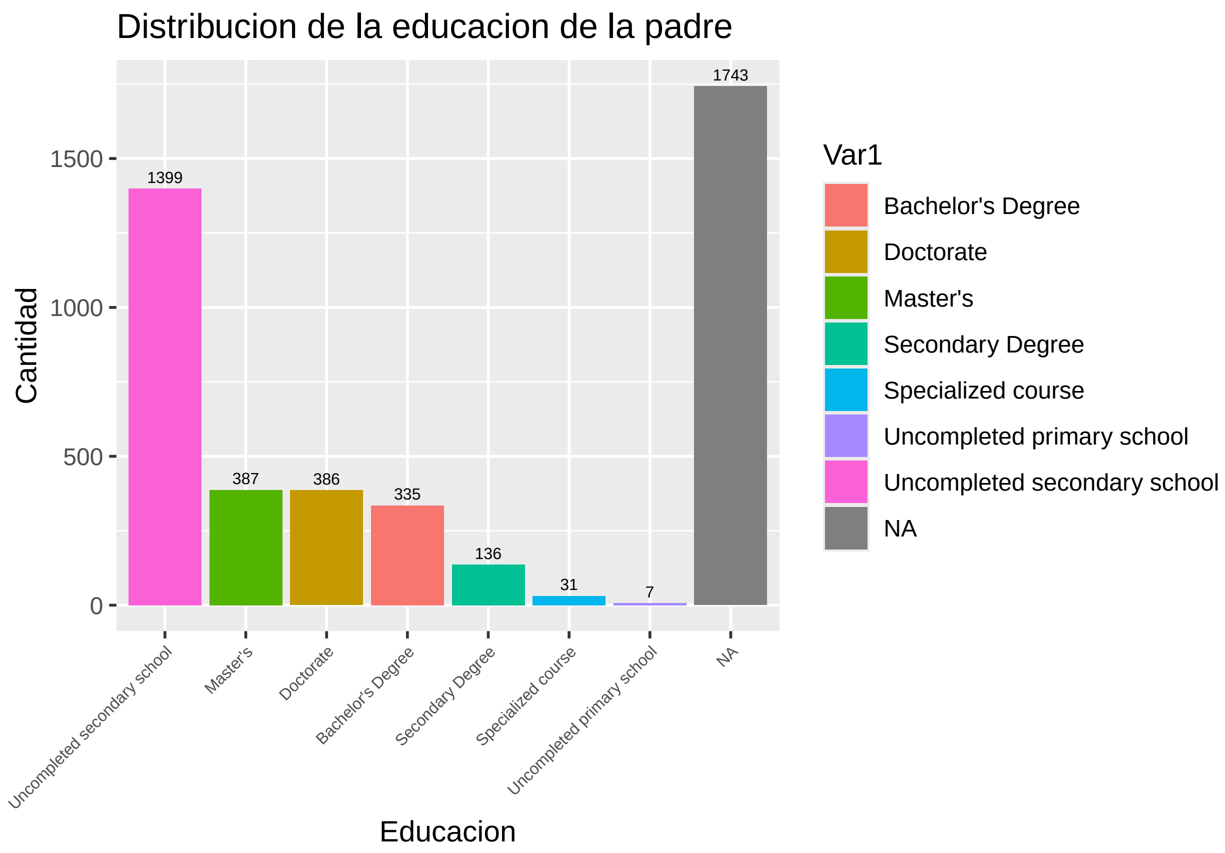
Ocupacion materna

```
contarMotherOccupation <- data.frame(table(datos$Mother.s.occupation.name, useNA = "always"))

ggplot(contarMotherOccupation, aes(x = reorder(Var1, -Freq), y = Freq, fill = Var1)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = Freq), vjust = -0.5, size = 2) +
  labs(x = "Educacion", y = "Cantidad", title = "Distribucion de la educacion de la madre") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 6))
```



Ocupacion paterna

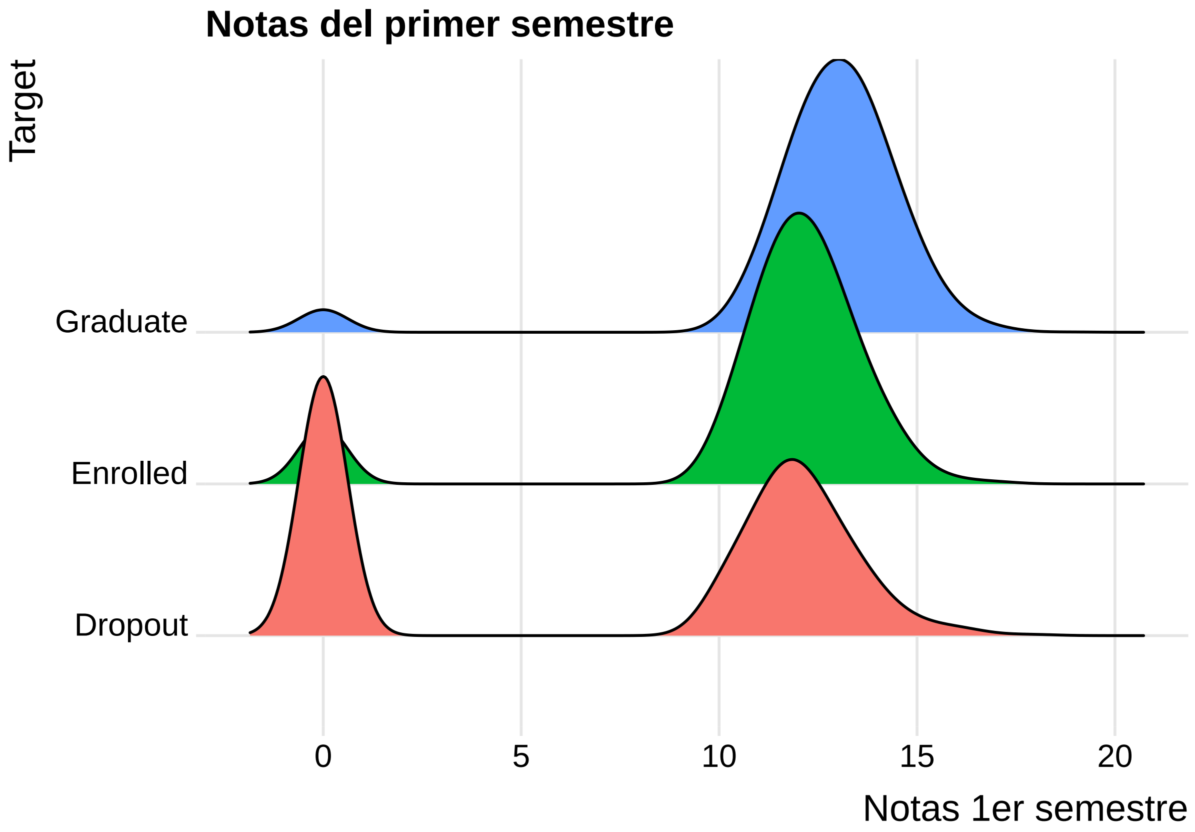
```
contarFatherOccupation <- data.frame(table(datos$Father.s.occupation.name, useNA = "always"))

ggplot(contarFatherOccupation, aes(x = reorder(Var1, -Freq), y = Freq, fill = Var1)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = Freq), vjust = -0.5, size = 2) +
  labs(x = "Educacion", y = "Cantidad", title = "Distribucion de la educacion de la padre") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 6))
```



Notas del 1er semestre según target

```r
ggplot(datos, aes(x = Curricular.units.1st.sem..grade., y = Target, fill = Target)) +
  geom_density_ridges() +
  theme_ridges() +
  theme(legend.position = "none") +
  labs(x = "Notas 1er semestre", title = "Notas del primer semestre")
```

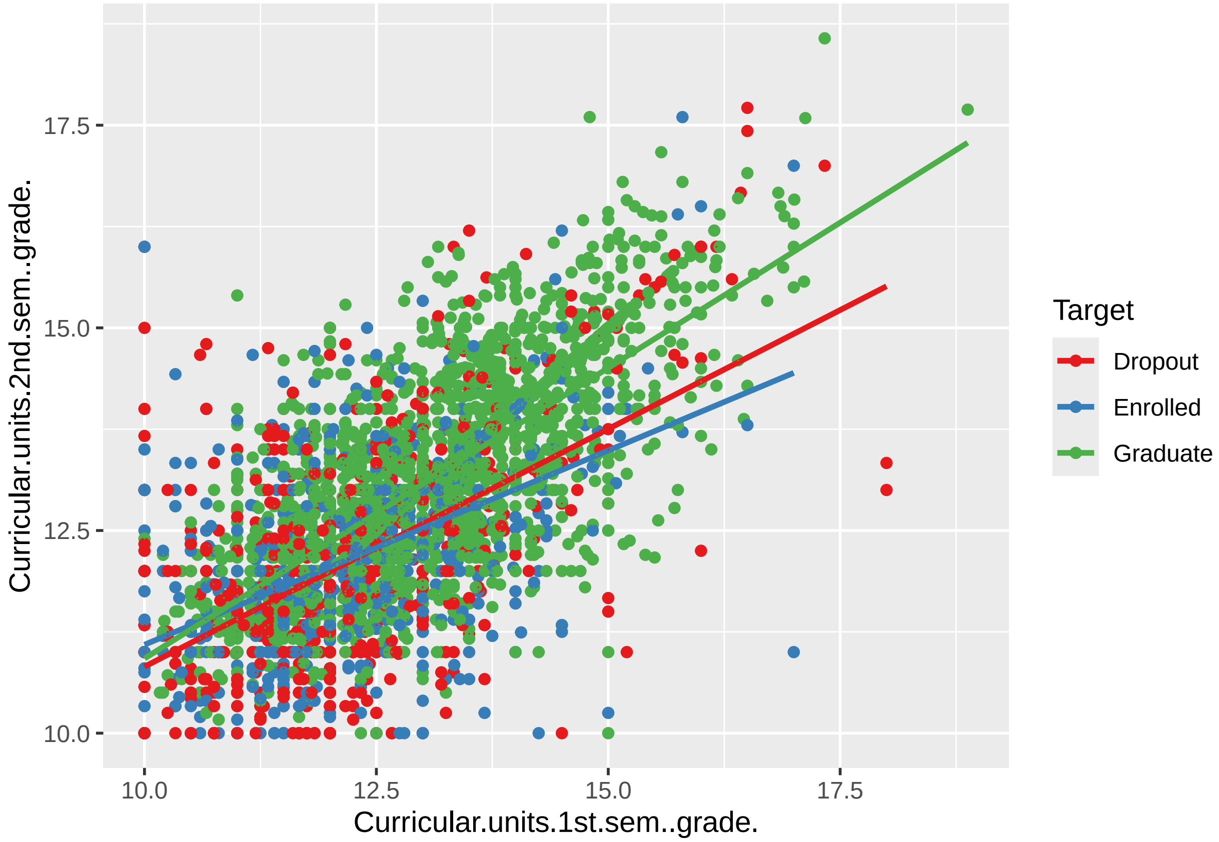
```
## Picking joint bandwidth of 0.616
```



Calificaciones del 1er y 2do Semestre por target y género

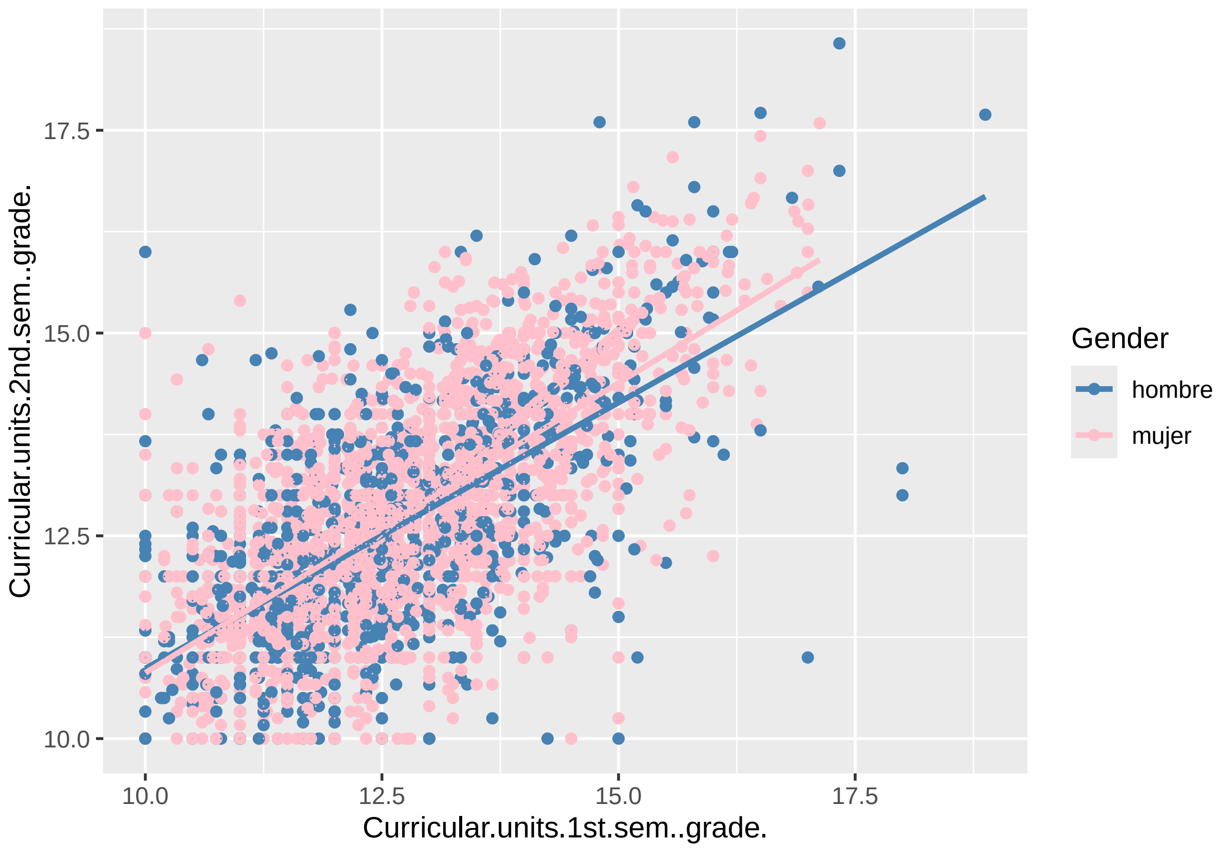
```
datos_notas_filtradas <- datos %>% filter(Curricular.units.1st.sem..grade. >= 10 & Curricular.units.2nd

ggplot(datos_notas_filtradas, aes(x=Curricular.units.1st.sem..grade., y=Curricular.units.2nd.sem..grade
  geom_smooth(method="lm", se=FALSE)+
  geom_point(size=0.05, shape=18) +
  scale_colour_brewer(palette = "Set1")
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



```
ggplot(datos_notas_filtradas, aes(x=Curricular.units.1st.sem..grade., y=Curricular.units.2nd.sem..grade
  geom_smooth(method="lm", se=FALSE)+
  geom_point(size=0.05, shape=18) +
  scale_color_manual(values = c("steelblue", "pink"))
```

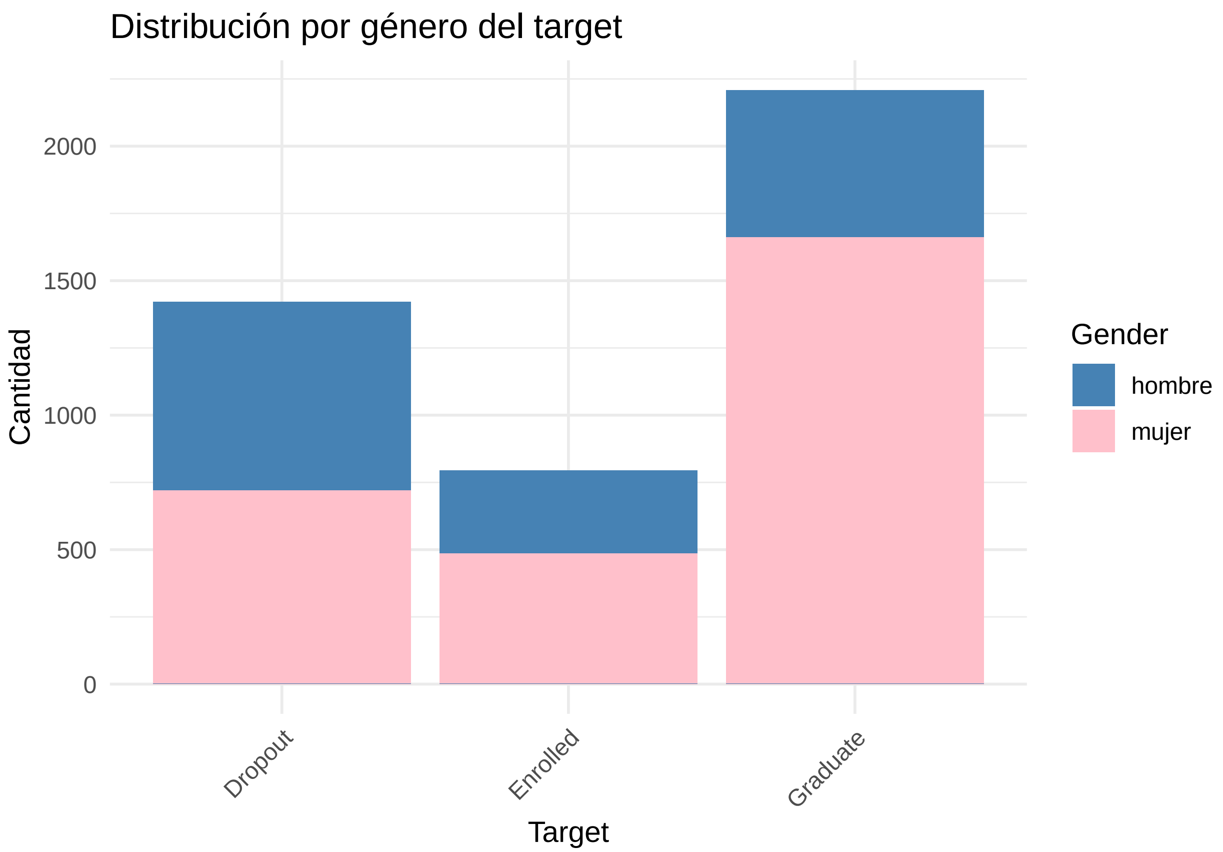
```
## `geom_smooth()` using formula = 'y ~ x'
```



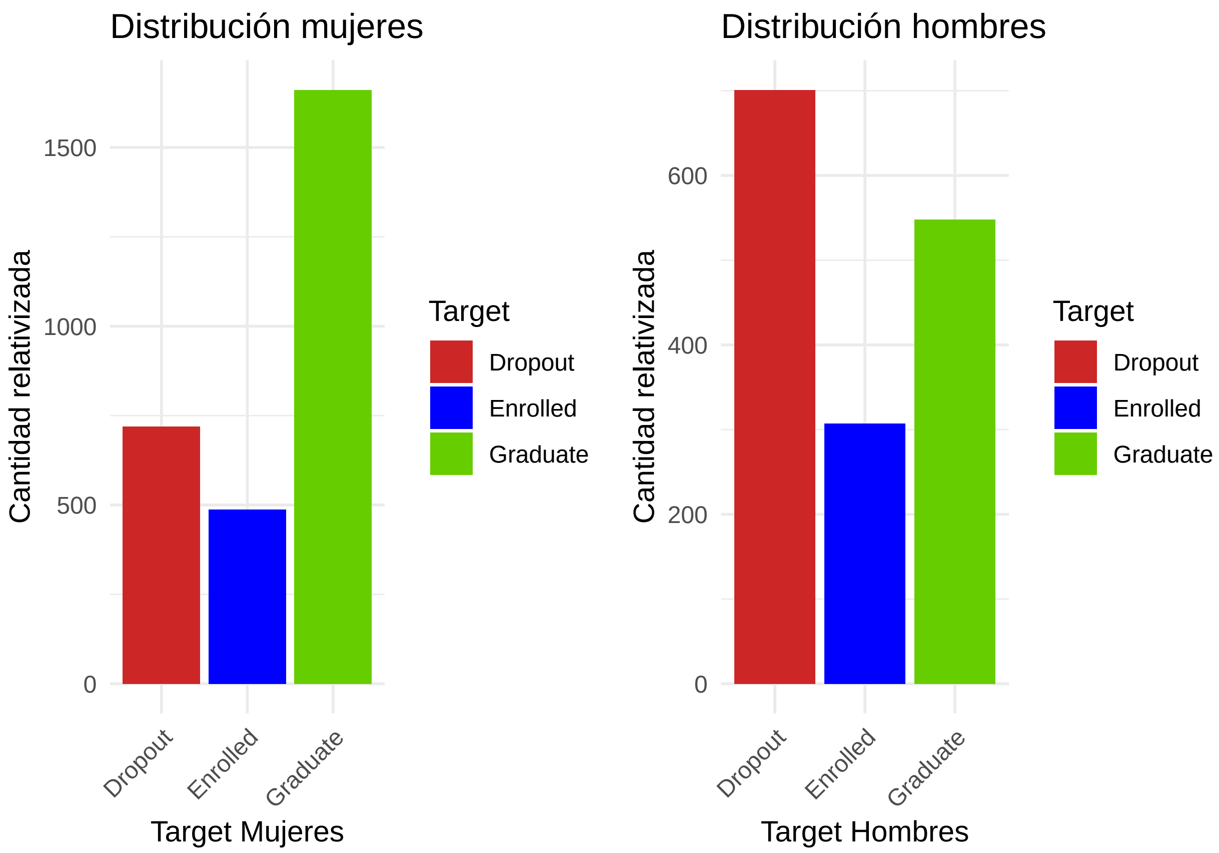
¿Hay relacion entre el género y el abandono de la carrera?

```
ggplot(data=datos, aes(x = Target, fill = Gender)) +
  geom_bar(position = "stack") +
  geom_bar(position = "stack", stat = "count", aes(y = ..prop..)) +
  scale_fill_manual(values = c("steelblue","pink")) +
  labs(x = "Target", y = "Cantidad", title = "Distribución por género del target") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: The dot-dot notation (‘..prop..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(prop)‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

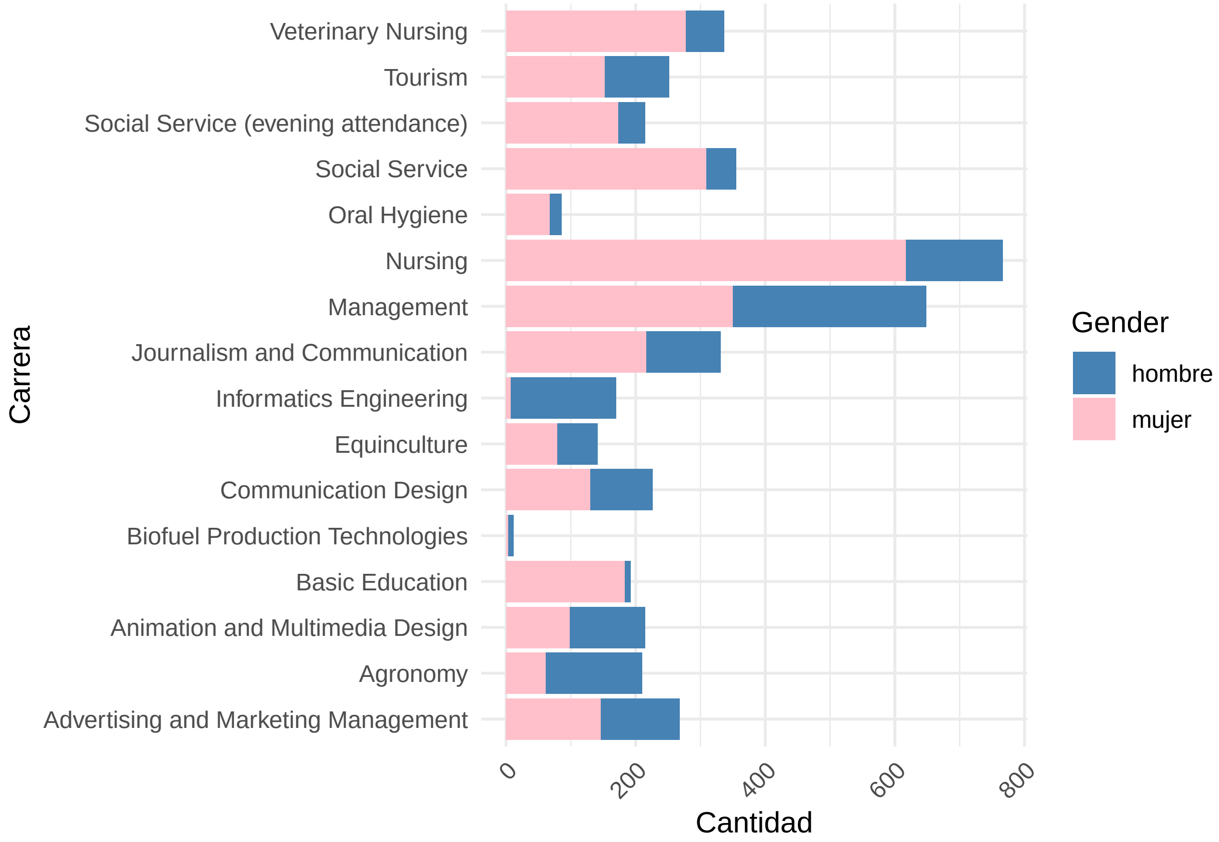


```
targetMujeres <- ggplot(data=datos_mujer, aes(x = Target, fill = Target)) +
  geom_bar(position = "stack") +
  scale_fill_manual(values = c("firebrick3", "blue", "chartreuse3")) +
  labs(x = "Target Mujeres", y = "Cantidad relativizada", title = "Distribución mujeres") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

targetHombres <- ggplot(data=datos_hombre, aes(x = Target, fill = Target)) +
  geom_bar(position = "stack") +
  scale_fill_manual(values = c("firebrick3", "blue", "chartreuse3")) +
  labs(x = "Target Hombres", y = "Cantidad relativizada", title = "Distribución hombres") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

grid.arrange(targetMujeres, targetHombres, ncol = 2)
```



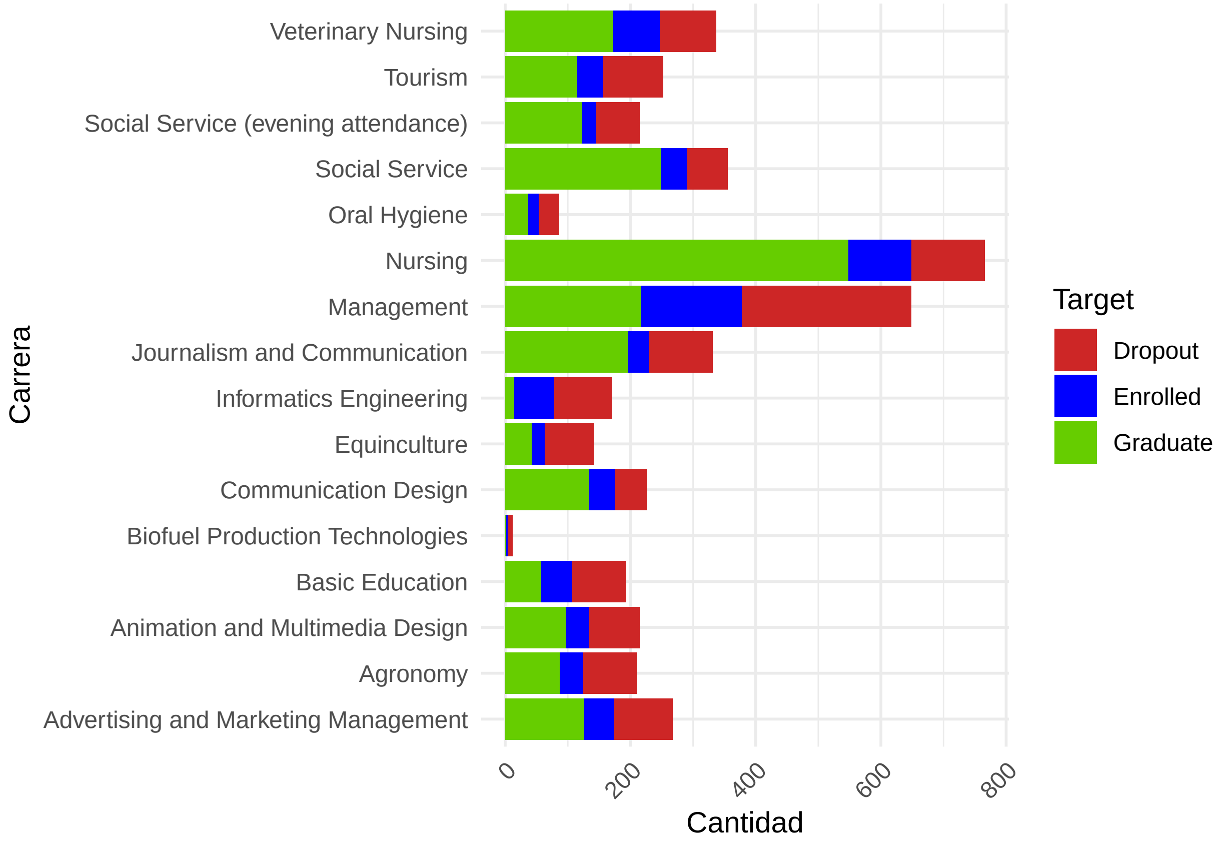
Carreras por género

```r
ggplot(data=datos, aes(x = Course, fill = Gender)) +
  geom_bar(position = "stack") +
  scale_fill_manual(values = c("steelblue", "pink")) +
  labs(x = "Carrera", y = "Cantidad") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  coord_flip()
```



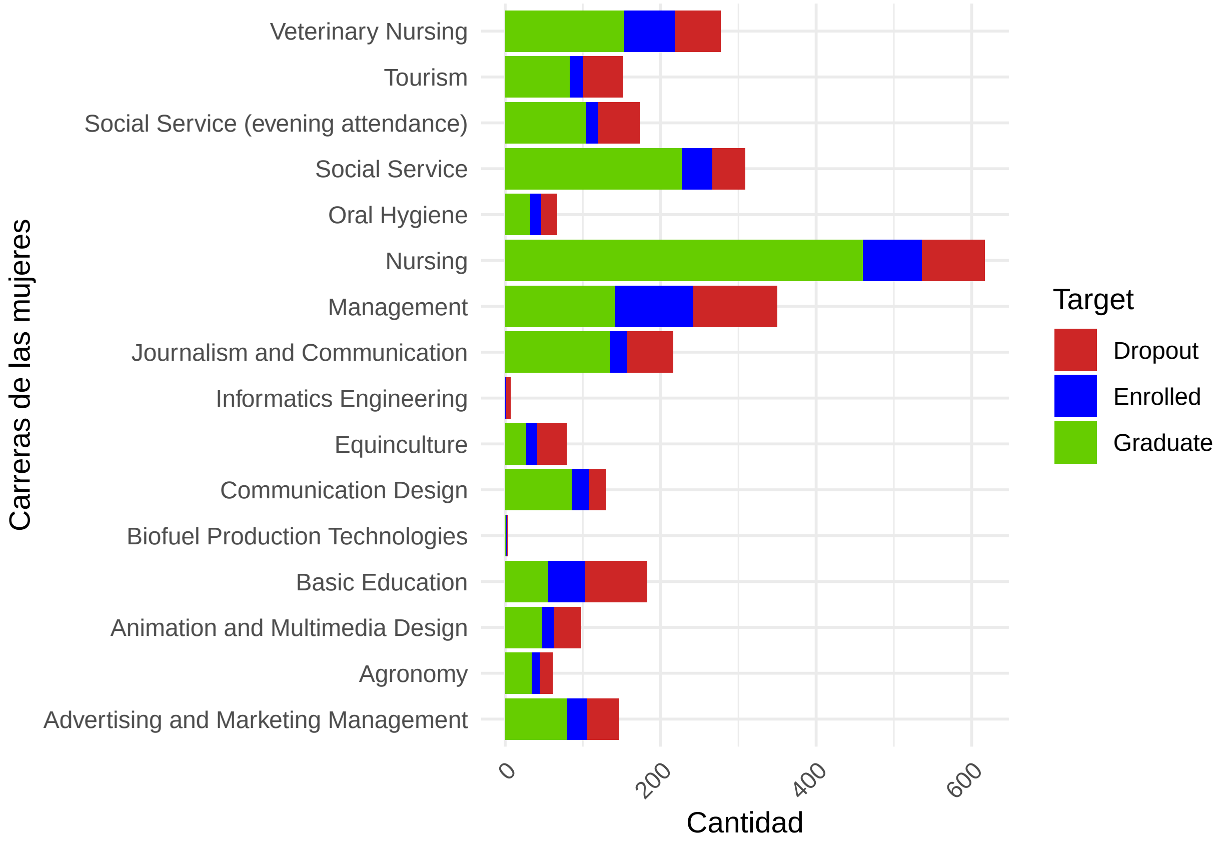
Carreras por target

```
ggplot(data=datos, aes(x = Course, fill = Target)) +
  geom_bar(position = "stack") +
  scale_fill_manual(values = c("firebrick3", "blue", "chartreuse3")) +
  labs(x = "Carrera", y = "Cantidad") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  coord_flip()
```

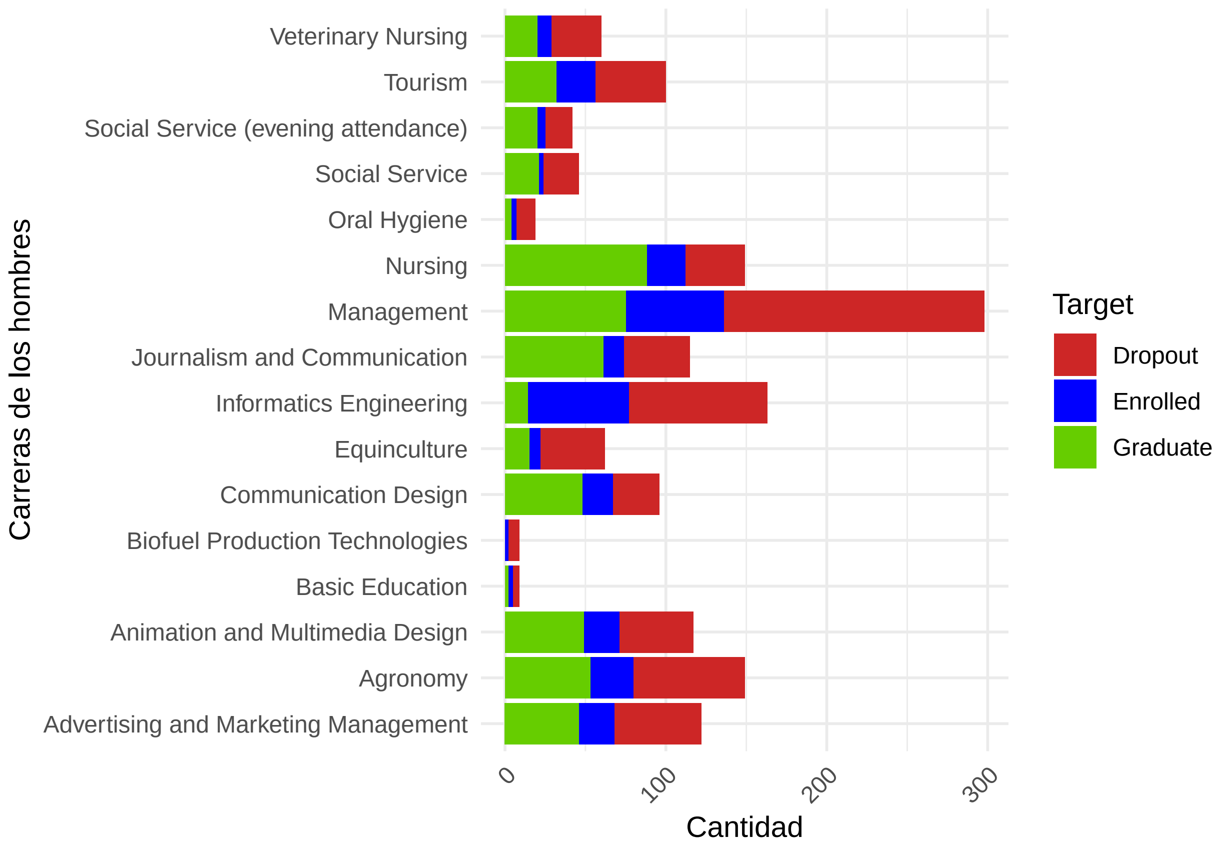


Carreras en mujeres y hombres por target

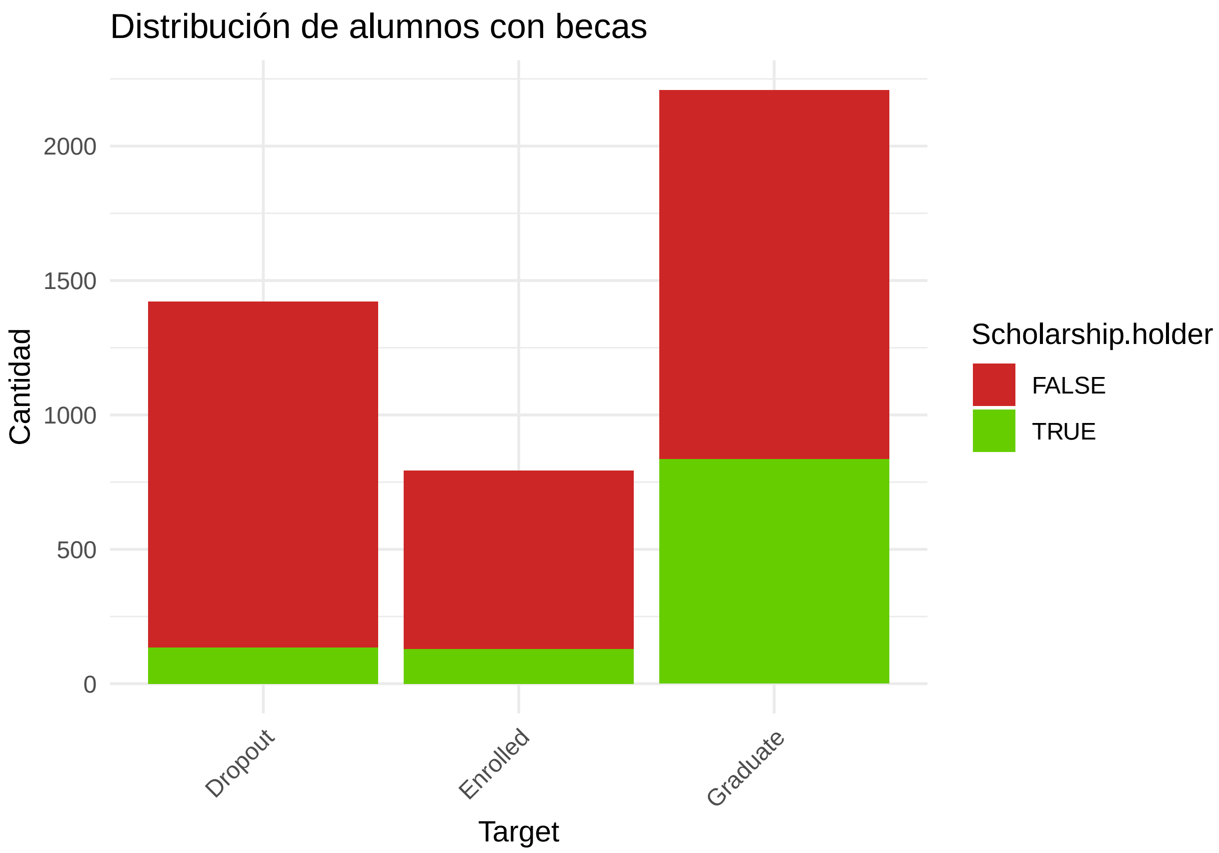
```
ggplot(data=datos_mujer, aes(x = Course, fill = Target)) +
  geom_bar(position = "stack") +
  scale_fill_manual(values = c("firebrick3", "blue", "chartreuse3")) +
  labs(x = "Carreras de las mujeres", y = "Cantidad") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  coord_flip()
```



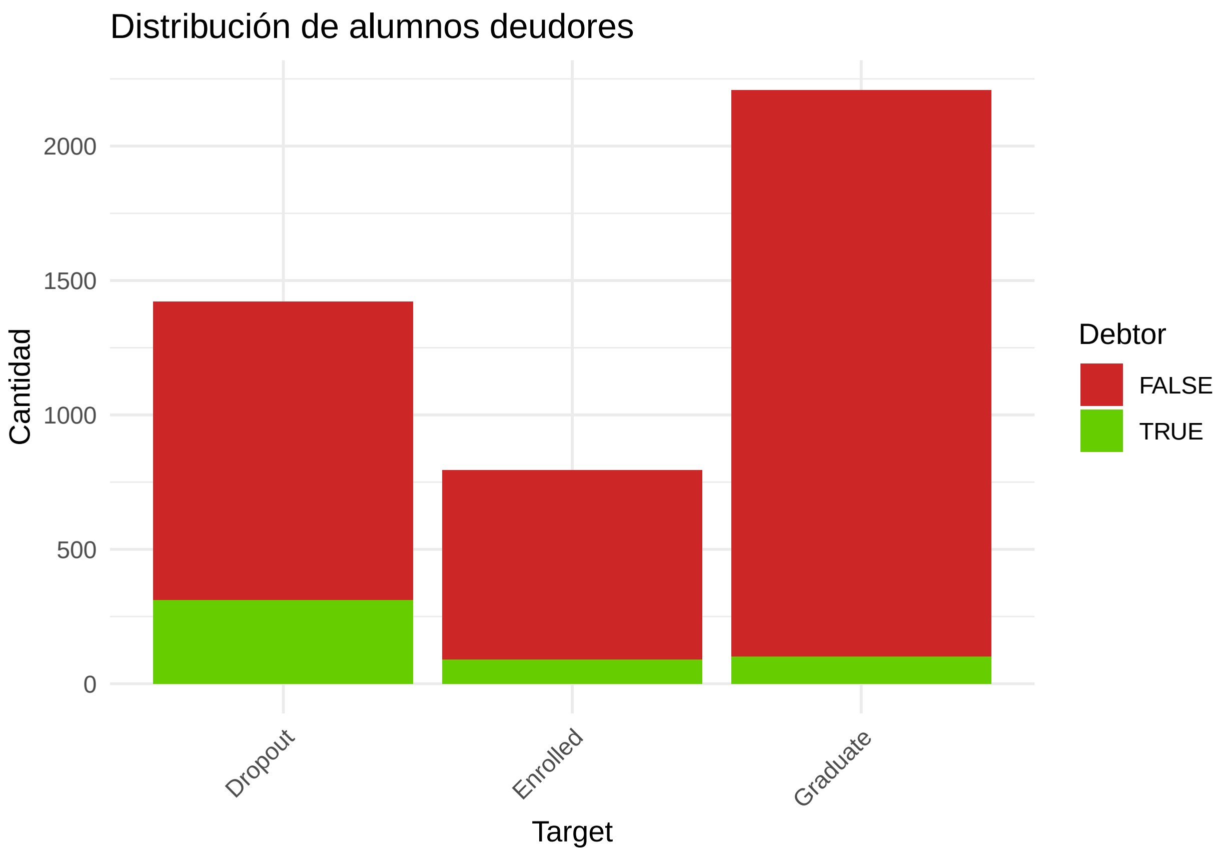
```
ggplot(data=datos_hombre, aes(x = Course, fill = Target)) +
  geom_bar(position = "stack") +
  scale_fill_manual(values = c("firebrick3", "blue", "chartreuse3")) +
  labs(x = "Carreras de los hombres", y = "Cantidad") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  coord_flip()
```



```
ggplot(data=datos, aes(x = Target, fill = Scholarship.holder)) +
  geom_bar(position = "stack") +
  scale_fill_manual(values = c("firebrick3","chartreuse3")) +
  labs(x = "Target", y = "Cantidad", title = "Distribución de alumnos con becas") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

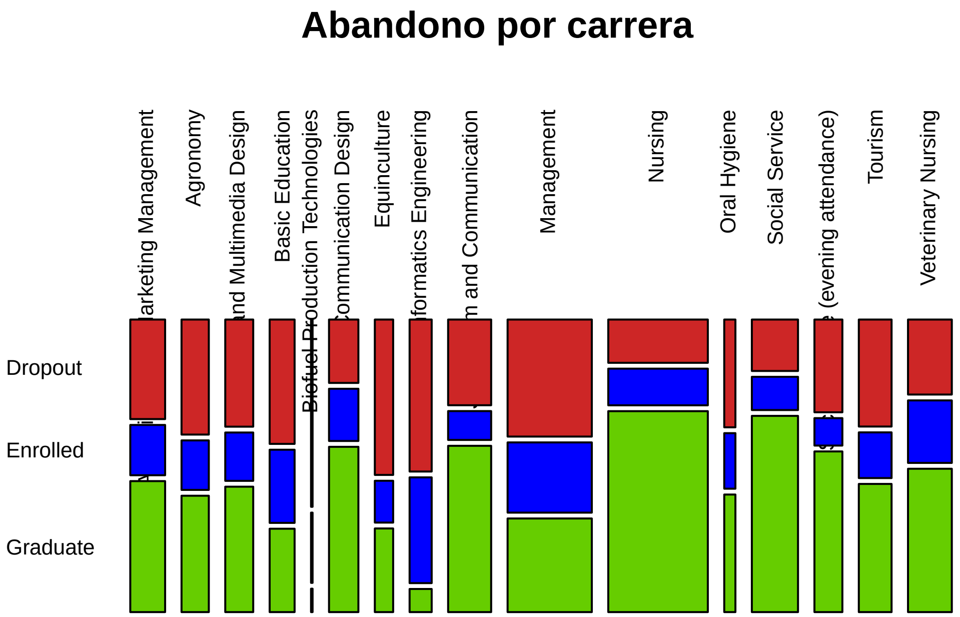


```r
ggplot(data=datos, aes(x = Target, fill = Debtor)) +
  geom_bar(position = "stack") +
  scale_fill_manual(values = c("firebrick3","chartreuse3")) +
  labs(x = "Target", y = "Cantidad", title = "Distribución de alumnos deudores") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



¿Qué carreras abandonan más?

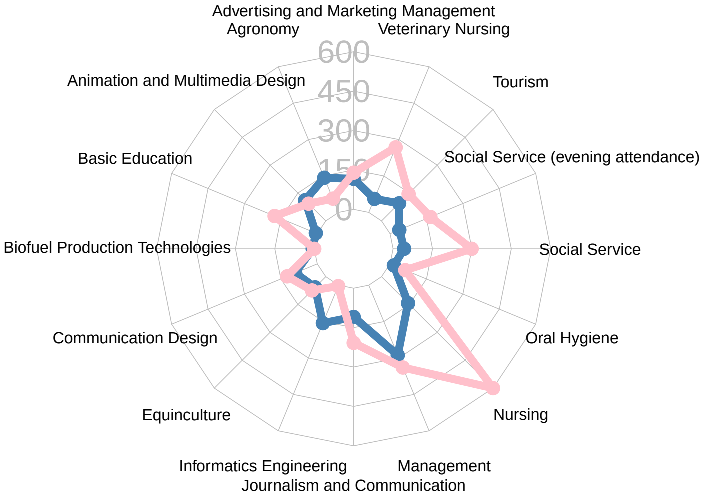
```
tablaCarrera = table(datos$Course,datos$Target)

par(las=2,cex.lab = 0.02)

mosaicplot(tablaCarrera, main="Abandono por carrera",
col=c("firebrick3", "blue", "chartreuse3"))
```



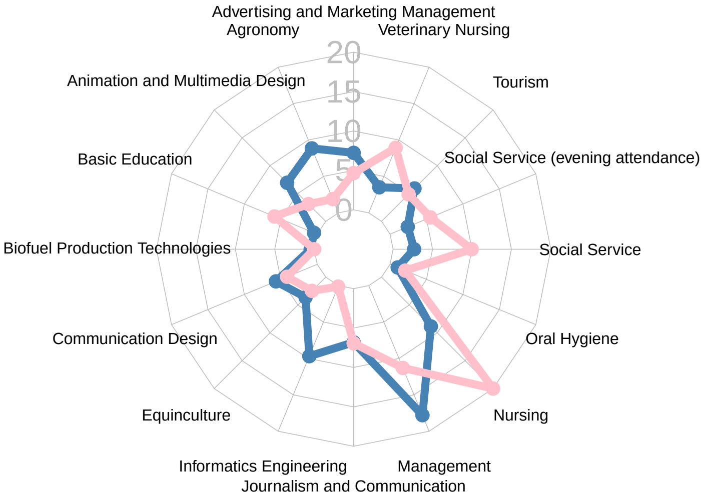
Género y carrera

```
table_df <- table(datos$Gender, datos$Course)

data <- as.data.frame.matrix(xtabs(Freq ~ Var1 + Var2, data = as.data.frame(table_df)))

datanorm <- as.data.frame(sapply(data, as.numeric))
datanorm <- t(apply(datanorm, 1, function(x) x/sum(x)))


rownames(data) <- c("hombre", "mujer")
rownames(datanorm) <- c("hombre", "mujer")

min_value <- rep(min(apply(data, 2, min)),length(unique(datos$Course)))
min_row <- data.frame(t(min_value))
colnames(min_row) <- colnames(data)
rownames(min_row) <- " "

max_value <- rep(max(apply(data, 2, max)),length(unique(datos$Course)))
max_row <- data.frame(t(max_value))
colnames(max_row) <- colnames(data)
rownames(max_row) <- " "

min_valuenorm<- rep(min(apply(datanorm, 2, min)),length(unique(datos$Course)))
```

```
min_rownorm <- data.frame(t(min_valuenorm))
colnames(min_rownorm) <- colnames(datanorm)
rownames(min_rownorm) <- " "

max_valuenorm <- rep(max(apply(datanorm, 2, max)),length(unique(datos$Course)))
max_rownorm <- data.frame(t(max_valuenorm))
colnames(max_rownorm) <- colnames(datanorm)
rownames(max_rownorm) <- " "

data <- rbind(max_row,min_row, data)
datanorm <- rbind(max_rownorm,min_rownorm, datanorm)

color_border <- c("steelblue","pink")


radarchart( data  , axistype=1 ,
    pcol=color_border , plwd=4 , plty=1,
    cglcol="grey", cglty=1, axislabcol="grey", caxislabels=seq(0,600,150), cglwd=0.4,
    vlcex=0.5
    )
```



```
#RELATIVIZADO EN PORCENTAJES HOMBRES/MUJERES
radarchart( datanorm  , axistype=1 ,
    pcol=color_border , plwd=4 , plty=1,
    cglcol="grey", cglty=1, axislabcol="grey", caxislabels=seq(0,20,5), cglwd=0.4,
    vlcex=0.5
    )
```


Advertising and Marketing Management
Agronomy
Veterinary Nursing
Tourism
Social Service (evening attendance)
Social Service
Oral Hygiene
Nursing
Management
Journalism and Communication
Informatics Engineering
Equinculture
Communication Design
Biofuel Production Technologies
Basic Education
Animation and Multimedia Design
20
15
10
5
0