1. Análisis Exploratorio

Pablo Parrilla Cañadas

2025-04-29

Librerías

```
library(dplyr)
```

```
##
## Adjuntando el paquete: 'dplyr'

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

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

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.3
```

```
library(reshape2)
library(gridExtra)
```

```
##
## Adjuntando el paquete: 'gridExtra'

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

```
library(skimr)
```

```
## Warning: package 'skimr' was built under R version 4.4.3
```

```
library(RColorBrewer)
library(scales)
library(corrplot)
```

```
## corrplot 0.94 loaded
```

```r
library(reshape2)
library(lsr)
```

```
## Warning: package 'lsr' was built under R version 4.4.3
```

```r
library(vcd)
```

```
## Warning: package 'vcd' was built under R version 4.4.3
```

```
## Cargando paquete requerido: grid
```

```r
library(rcompanion)
```

```
## Warning: package 'rcompanion' was built under R version 4.4.3
```

```r
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.4.2
```

```
## Cargando paquete requerido: lattice
```

```r
library(factoextra)
```

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

```r
library(lubridate)
```

```
##
## Adjuntando el paquete: 'lubridate'
```

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

```r
library(tidyr)
```

```
##
## Adjuntando el paquete: 'tidyr'
```

```
## The following object is masked from 'package:reshape2':
##
##     smiths
```

```r
library(forcats)
library(FactoMineR)
library(fastDummies)
```

```
## Warning: package 'fastDummies' was built under R version 4.4.2
```

```r
load("../../Datos/Capítulos/Preprocesado.RData")
```

Datos sociodemográficos

```r
a=skim(sociodemografia)
a=as.data.frame(a)
```

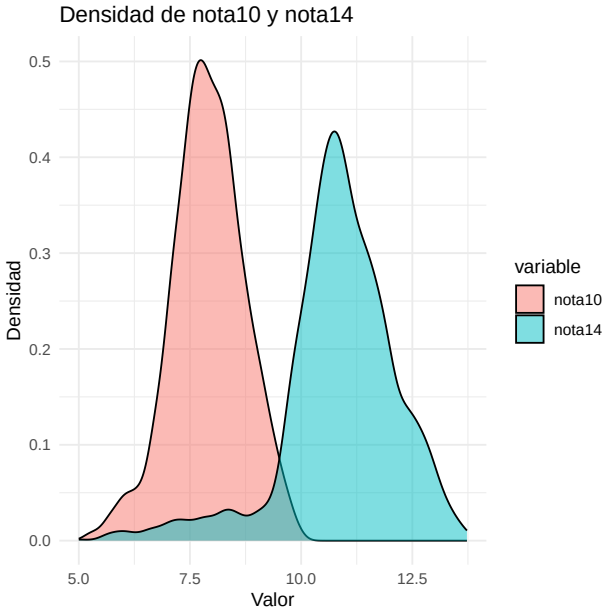
Gráficas de las numéricas

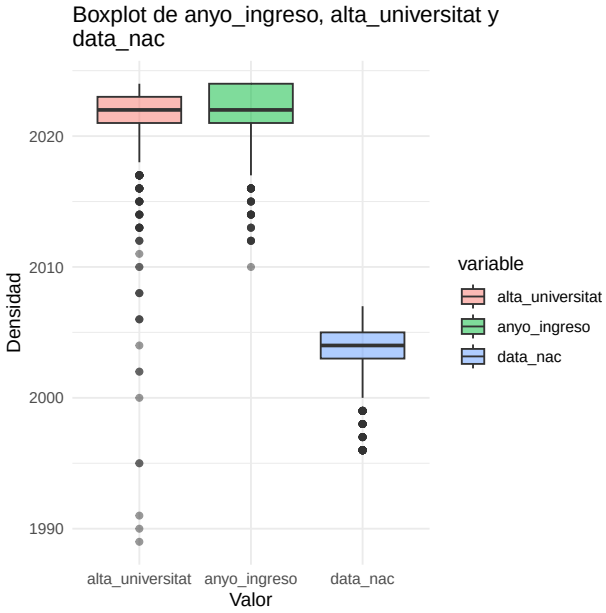
```r
df_long=sociodemografia %>%
  pivot_longer(cols = c(nota10, nota14), names_to = "variable", values_to = "valor")

a=ggplot(df_long, aes(x = valor, fill = variable)) +
  geom_density(alpha = 0.5) +
  labs(title = "Densidad de nota10 y nota14", x = "Valor", y = "Densidad") +
  theme_minimal()
```

```r
df_long=sociodemografia %>%
  pivot_longer(cols = c(anyo_ingreso, alta_universitat,data_nac), names_to = "variable", values_to = "va

b=ggplot(df_long, aes(x = variable,y=valor, fill = variable)) +
  geom_boxplot(alpha = 0.5) +
  labs(title = "Boxplot de anyo_ingreso, alta_universitat y \ndata_nac", x = "Valor", y = "Densidad") +
  theme_minimal()

grid.arrange(a, b, ncol = 2)
```

```
## Warning: Removed 66 rows containing non-finite outside the scale range
## (`stat_density()`).
```

```
## Warning: Removed 5 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



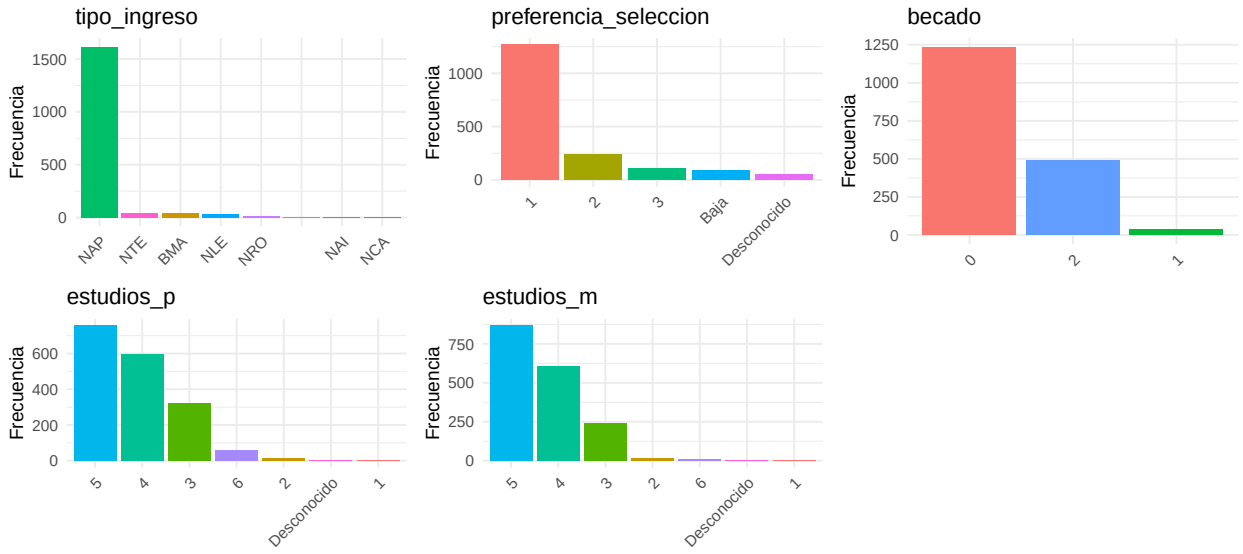


Cambiando los NA a desconocido en las categóricas.

```
sociodemografia[sapply(sociodemografia, is.factor)] <-
  lapply(sociodemografia[sapply(sociodemografia, is.factor)], function(x) {
    x=addNA(x)
    levels(x)[is.na(levels(x))]="Desconocido"
    x[is.na(x)]="Desconocido"
    return(x)})
```

Graficar categóricas.

```
multi_vars=c("tipo_ingreso", "preferencia_seleccion","becado", "estudios_p", "estudios_m")

plots_multi=lapply(multi_vars, function(var) {
  ggplot(sociodemografia, aes(x = fct_infreq(.data[[var]]), fill = .data[[var]])) +
    geom_bar() +
    scale_fill_discrete() +
    labs(title = var, x = NULL, y = "Frecuencia") +
    theme_minimal() +
    theme(legend.position = "none", axis.text.x = element_text(angle = 45, hjust = 1))
})

grid.arrange(grobs = plots_multi, ncol = 3)
```




```
table(sociodemografia$estudios_m)
```

```
##
##           1           2           3           4           5           6
##           3          18         244         606         867          11
## Desconocido
##           5
```

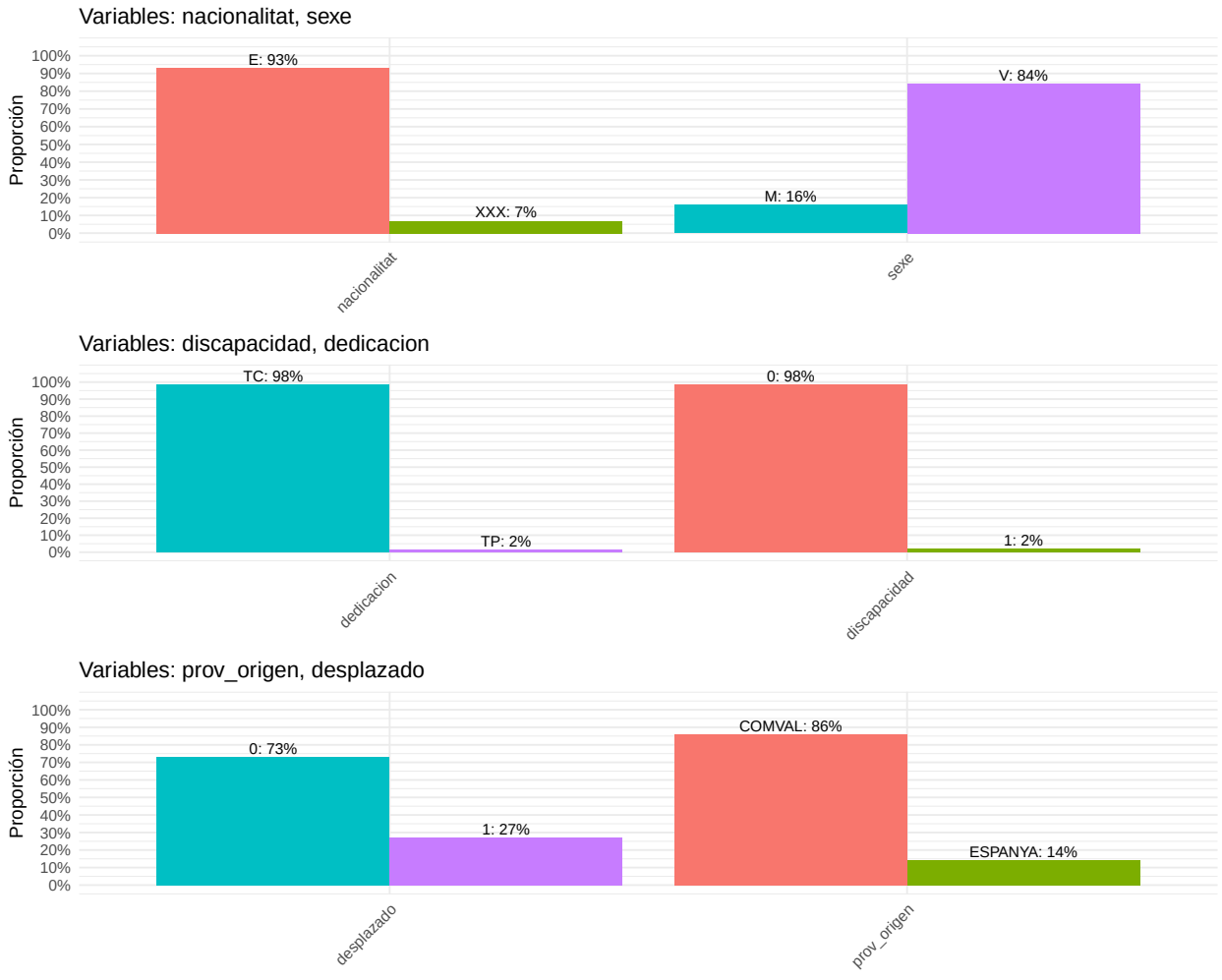
```
cat("")
table(sociodemografia$estudios_p)
```

```
##
##           1           2           3           4           5           6
##           4          15         323         594         756          57
## Desconocido
##           5
```

Graficar binarias.

```
binarias_1=c("nacionalitat", "sexe")
binarias_3=c("prov_origen", "desplazado")
binarias_2=c("discapacidad","dedicacion" )

crear_plot_binarias=function(vars) {
  df=sociodemografia %>%
    select(all_of(vars)) %>%
    pivot_longer(cols = everything(), names_to = "Variable", values_to = "Valor") %>%
    filter(!is.na(Valor)) %>%
    group_by(Variable, Valor) %>%
    summarise(n = n(), .groups = "drop") %>%
    group_by(Variable) %>%
    mutate(Proporcion = n / sum(n),
           Etiqueta = paste0(Valor, ": ", round(Proporcion * 100), "%"))
```

```r
  ggplot(df, aes(x = Variable, y = Proporcion, fill = Valor)) +
    geom_col(position = position_dodge(width = 0.9)) +
    geom_text(aes(label = Etiqueta), position = position_dodge(width = 0.9),
              vjust = -0.3, size = 3) +
    scale_y_continuous(labels = scales::percent_format(accuracy = 1),
                       breaks = seq(0, 1, 0.1), limits = c(0, 1.05)) +
    labs(title = paste("Variables:", paste(vars, collapse = ", ")),
         y = "Proporción", x = NULL) +
    theme_minimal() +
    theme(legend.position = "none",
          axis.text.x = element_text(angle = 45, hjust = 1))}


plot1=crear_plot_binarias(binarias_1)
plot2=crear_plot_binarias(binarias_2)
plot3=crear_plot_binarias(binarias_3)


grid.arrange(plot1, plot2, plot3, nrow = 3)
```



Datos académicos

```
summary(academicas)
```

```
##    dni_hash         curso_mas_bajo curso_mas_alto   cred_mat1
##  Length:1754        1:596          1:436          Min.   : 0.00
##  Class :character   2:552          2:409          1st Qu.: 0.00
##  Mode  :character   3:349          3:399          Median : 0.00
##                     4:257          4:510          Mean   :16.09
##                                                   3rd Qu.:36.00
##                                                   Max.   :60.00
##
##    cred_mat2       cred_mat3       cred_mat4     cred_sup_normal
##  Min.   : 0.00   Min.   : 0.00   Min.   : 0.00   Min.   : 0.000
##  1st Qu.: 0.00   1st Qu.: 0.00   1st Qu.: 0.00   1st Qu.: 0.000
##  Median : 0.00   Median : 0.00   Median : 0.00   Median : 0.000
##  Mean   :15.37   Mean   :13.28   Mean   :10.69   Mean   : 1.308
##  3rd Qu.:30.00   3rd Qu.:19.50   3rd Qu.:13.12   3rd Qu.: 0.000
##  Max.   :64.50   Max.   :82.50   Max.   :73.50   Max.   :30.000
##
##  cred_sup_espec       cred_sup       cred_mat_normal  cred_mat_movilidad
##  Min.   :  0.000   Min.   :  0.000   Min.   :  0.00   Min.   : 0.0000
##  1st Qu.:  0.000   1st Qu.:  0.000   1st Qu.: 57.00   1st Qu.: 0.0000
##  Median :  0.000   Median :  0.000   Median : 60.00   Median : 0.0000
##  Mean   :  1.329   Mean   :  2.636   Mean   : 55.84   Mean   : 0.8586
##  3rd Qu.:  0.000   3rd Qu.:  0.000   3rd Qu.: 60.50   3rd Qu.: 0.0000
##  Max.   :126.000   Max.   :126.000   Max.   :114.00   Max.   :66.0000
##
##  cred_ptes_acta  cred_mat_practicas cred_mat_sem_a  cred_mat_sem_b
##  Min.   : 0.00   Min.   : 0.0       Min.   : 4.50   Min.   : 0.00
##  1st Qu.:39.00   1st Qu.: 0.0       1st Qu.:21.00   1st Qu.:21.00
##  Median :60.00   Median : 0.0       Median :30.00   Median :25.50
##  Mean   :49.06   Mean   : 2.6       Mean   :27.28   Mean   :23.65
##  3rd Qu.:60.00   3rd Qu.: 0.0       3rd Qu.:30.00   3rd Qu.:30.00
##  Max.   :78.00   Max.   :18.0       Max.   :57.00   Max.   :55.50
##
##   cred_mat_anu  cred_mat_total  cred_sup_sem_a    cred_sup_sem_b
##  Min.   : 0.0   Min.   : 4.50   Min.   : 0.0000   Min.   : 0.00000
##  1st Qu.: 0.0   1st Qu.:52.50   1st Qu.: 0.0000   1st Qu.: 0.00000
##  Median : 0.0   Median :60.00   Median : 0.0000   Median : 0.00000
##  Mean   : 4.5   Mean   :55.43   Mean   : 0.6209   Mean   : 0.01368
##  3rd Qu.: 9.0   3rd Qu.:60.00   3rd Qu.: 0.0000   3rd Qu.: 0.00000
##  Max.   :18.0   Max.   :96.00   Max.   :30.0000   Max.   :12.00000
##
##  cred_sup_total    rendimiento_cuat_a rendimiento_total exento_npp
##  Min.   : 0.0000   Min.   :  0.000    Min.   :  0.000   0:1338
##  1st Qu.: 0.0000   1st Qu.:  0.000    1st Qu.:  0.000   1: 416
##  Median : 0.0000   Median :  0.000    Median :  0.000
##  Mean   : 0.6345   Mean   :  1.979    Mean   :  1.313
##  3rd Qu.: 0.0000   3rd Qu.:  0.000    3rd Qu.:  0.000
##  Max.   :30.0000   Max.   :100.000    Max.   :100.000
##                    NA's   :54         NA's   :54
```

```
##  anyo_inicio_estudios es_retitulado es_adaptado  cred_sup_1o     cred_sup_2o
##  Min.   :2010         0:1752        0:1748      Min.   : 0.00   Min.   : 0.00
##  1st Qu.:2021         1:   2        1:   6      1st Qu.:18.00   1st Qu.: 0.00
##  Median :2022                                   Median :60.00   Median :18.00
##  Mean   :2022                                   Mean   :43.65   Mean   :28.16
##  3rd Qu.:2024                                   3rd Qu.:60.00   3rd Qu.:60.00
##  Max.   :2024                                   Max.   :60.00   Max.   :64.50
##
##   cred_sup_3o     cred_sup_4o       practicas       actividades
##  Min.   : 0.00   Min.   : 0.000   Min.   : 0.000   Min.   :0.200
##  1st Qu.: 0.00   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.:1.000
##  Median : 0.00   Median : 0.000   Median : 0.000   Median :4.400
##  Mean   :15.52   Mean   : 2.428   Mean   : 1.656   Mean   :3.989
##  3rd Qu.:37.50   3rd Qu.: 0.000   3rd Qu.: 0.000   3rd Qu.:5.770
##  Max.   :82.50   Max.   :60.000   Max.   :18.000   Max.   :9.000
##                                                    NA's   :1345
##      ajuste        cred_sup_tit    cred_pend_sup_tit     asig1
##  Min.   : 7.500   Min.   :  0.00   Min.   :  0.0     Min.   : 0.0000
##  1st Qu.: 7.500   1st Qu.: 24.00   1st Qu.: 73.5     1st Qu.: 0.0000
##  Median : 7.500   Median : 79.50   Median :160.5     Median : 0.0000
##  Mean   : 9.461   Mean   : 92.79   Mean   :147.3     Mean   : 0.6345
##  3rd Qu.: 9.000   3rd Qu.:166.50   3rd Qu.:216.0     3rd Qu.: 0.0000
##  Max.   :30.000   Max.   :244.50   Max.   :240.0     Max.   :30.0000
##  NA's   :1702
##      pract1           activ1          total1      rend_total_ultimo
##  Min.   : 0.500   Min.   :0.200   Min.   : 4.80   Min.   :  0.00
##  1st Qu.: 4.000   1st Qu.:1.000   1st Qu.: 9.50   1st Qu.: 70.59
##  Median : 6.000   Median :3.500   Median :12.91   Median :100.00
##  Mean   : 7.309   Mean   :3.494   Mean   :12.87   Mean   : 82.04
##  3rd Qu.: 9.500   3rd Qu.:5.200   3rd Qu.:14.50   3rd Qu.:100.00
##  Max.   :18.000   Max.   :9.000   Max.   :27.92   Max.   :100.00
##  NA's   :1707     NA's   :1512    NA's   :1733    NA's   :445
##  rend_total_penultimo rend_total_antepenultimo    abandono
##  Min.   :  0.00       Min.   :  0.00           Min.   :0.00000
##  1st Qu.: 69.13       1st Qu.:  0.00           1st Qu.:0.00000
##  Median :100.00       Median :  0.00           Median :0.00000
##  Mean   : 80.83       Mean   : 25.07           Mean   :0.02965
##  3rd Qu.:100.00       3rd Qu.: 54.54           3rd Qu.:0.00000
##  Max.   :100.00       Max.   :100.00           Max.   :1.00000
##  NA's   :822
```

```
a=skim(academicas)
```

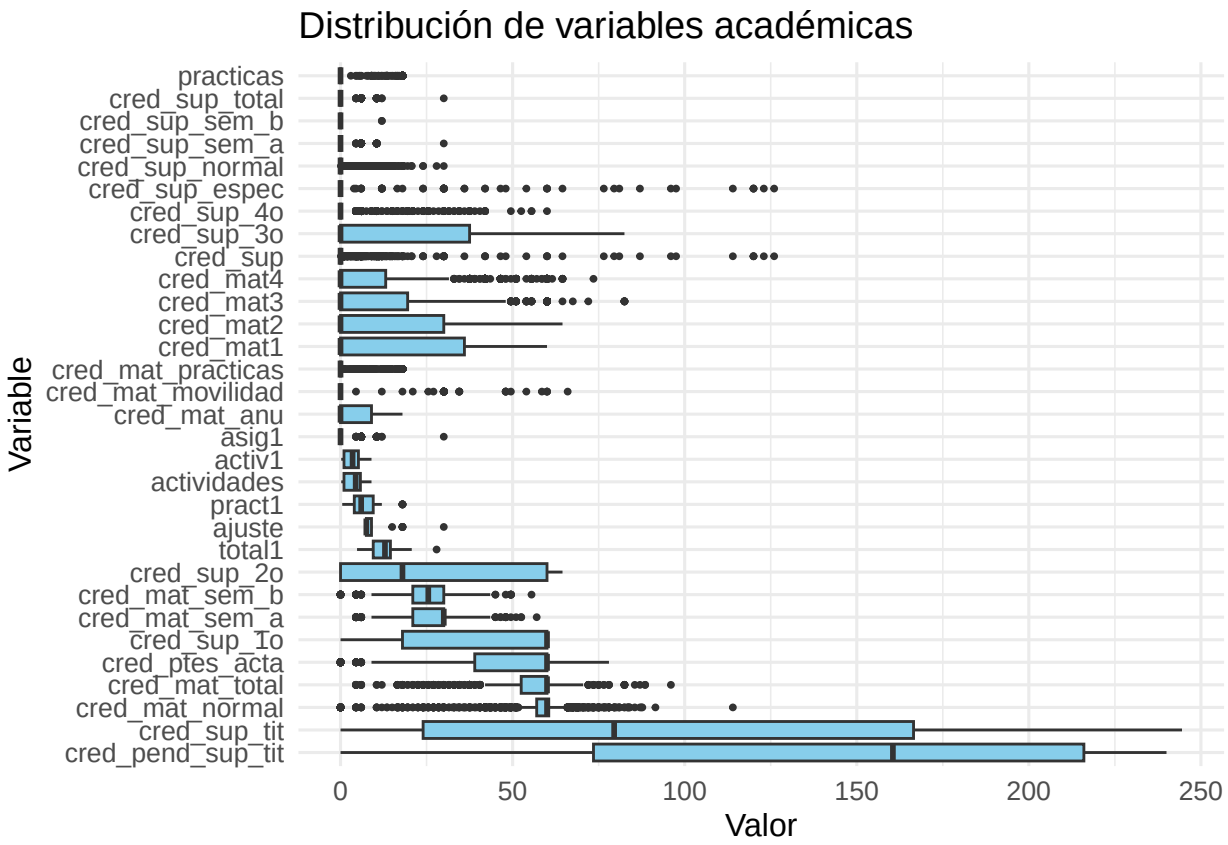
Graficar numéricas.

```
vars_academicas=c(
  "cred_mat1", "cred_mat2", "cred_mat3", "cred_mat4", "cred_sup_normal", "cred_sup_espec", "cred_sup",
  "cred_mat_normal", "cred_mat_movilidad", "cred_ptes_acta", "cred_mat_practicas", "cred_mat_sem_a",
  "cred_mat_sem_b", "cred_mat_anu", "cred_mat_total", "cred_sup_sem_a", "cred_sup_sem_b",
  "cred_sup_total", "cred_sup_1o", "cred_sup_2o", "cred_sup_3o", "cred_sup_4o", "practicas", "actividade
  "ajuste", "cred_sup_tit", "cred_pend_sup_tit", "asig1", "pract1", "activ1", "total1"
)
```

```
df_long=academicas %>%
  select(all_of(vars_academicas)) %>%
  pivot_longer(cols = everything(), names_to = "Variable", values_to = "Valor")


ggplot(df_long, aes(x = Valor, y = fct_reorder(Variable, -Valor, .fun = median, na.rm = TRUE))) +
  geom_boxplot(fill = "skyblue", outlier.size = 0.7) +
  labs(title = "Distribución de variables académicas", x = "Valor", y = "Variable") +
  theme_minimal(base_size = 12)
```

```
## Warning: `fct_reorder()` removing 7999 missing values.
## i Use `.na_rm = TRUE` to silence this message.
## i Use `.na_rm = FALSE` to preserve NAs.
```

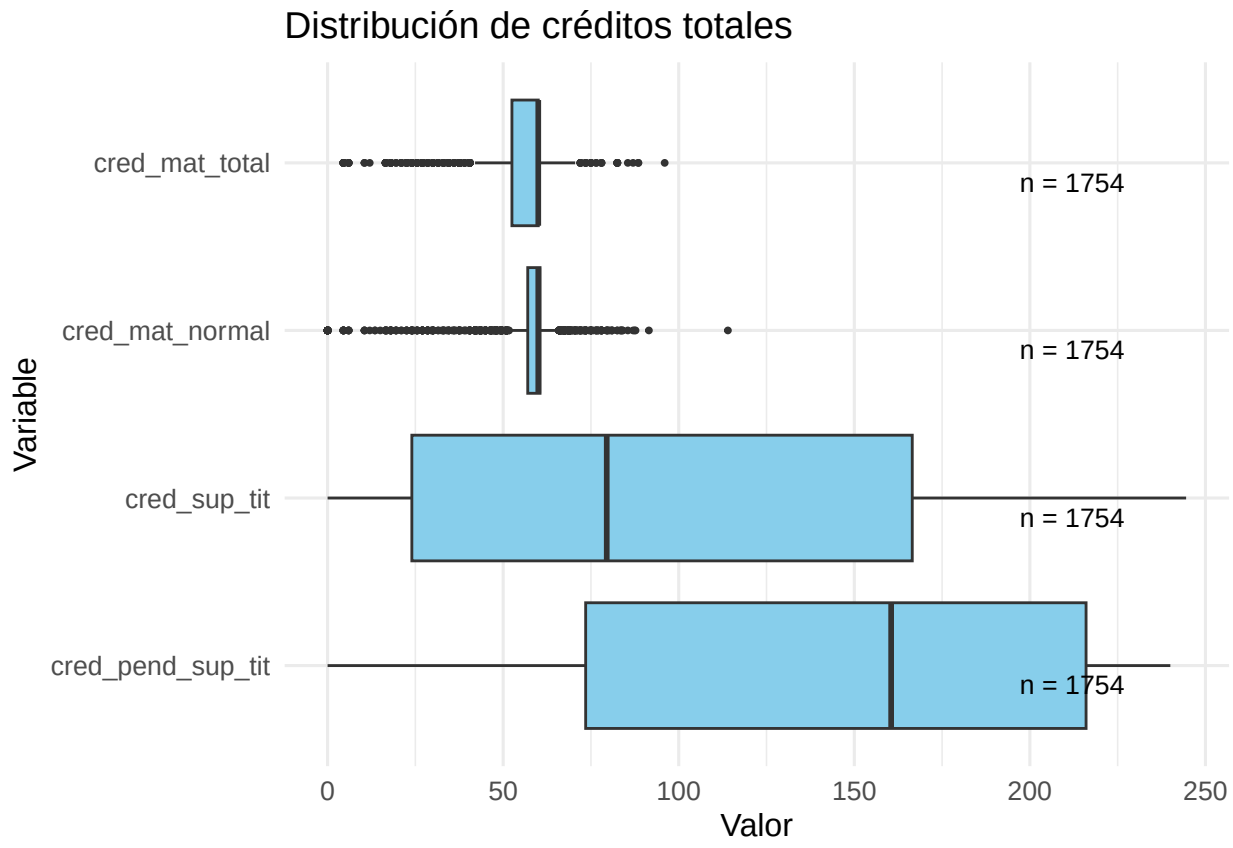
```
## Warning: Removed 7999 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



No se ve mucho así que las dividimos.

```
grupo_creditos_altos=c("cred_pend_sup_tit", "cred_sup_tit", "cred_mat_normal", "cred_mat_total")

grupo_creditos_moderados=c( "cred_ptes_acta", "cred_sup_1o", "cred_sup_2o", "cred_sup_3o","cred_sup_4o"

grupo_creditos_individuales=c("cred_mat1", "cred_mat2", "cred_mat3", "cred_mat4", "cred_mat_sem_a",  "cr
```

```r
grupo_creditos_especiales=c("cred_sup", "cred_sup_espec", "cred_sup_normal", "cred_sup_sem_a", "cred_su

grupo_otros_academicos=c("practicas", "pract1", "asig1", "activ1", "actividades", "ajuste", "total1")

rendimientos=c("rendimiento_cuat_a", "rendimiento_total", "rend_total_ultimo", "rend_total_penultimo","



df_long=academicas %>%
  select(all_of(grupo_creditos_altos)) %>%
  pivot_longer(cols = everything(), names_to = "Variable", values_to = "Valor")

conteos=df_long %>%
  filter(!is.na(Valor)) %>%
  group_by(Variable) %>%
  summarise(n = n())


ggplot(df_long, aes(x = Valor, y = fct_reorder(Variable, -Valor, .fun = median, na.rm = TRUE))) +
  geom_boxplot(fill = "skyblue", outlier.size = 0.7) +
  geom_text(data = conteos, aes(x = Inf, y = fct_reorder(Variable, -n), label = paste0("n = ", n)),
            hjust = 2,  size = 3.5, inherit.aes = FALSE,vjust = 1.5) +
  labs(title = "Distribución de créditos totales", x = "Valor", y = "Variable") +
  theme_minimal(base_size = 12)
```

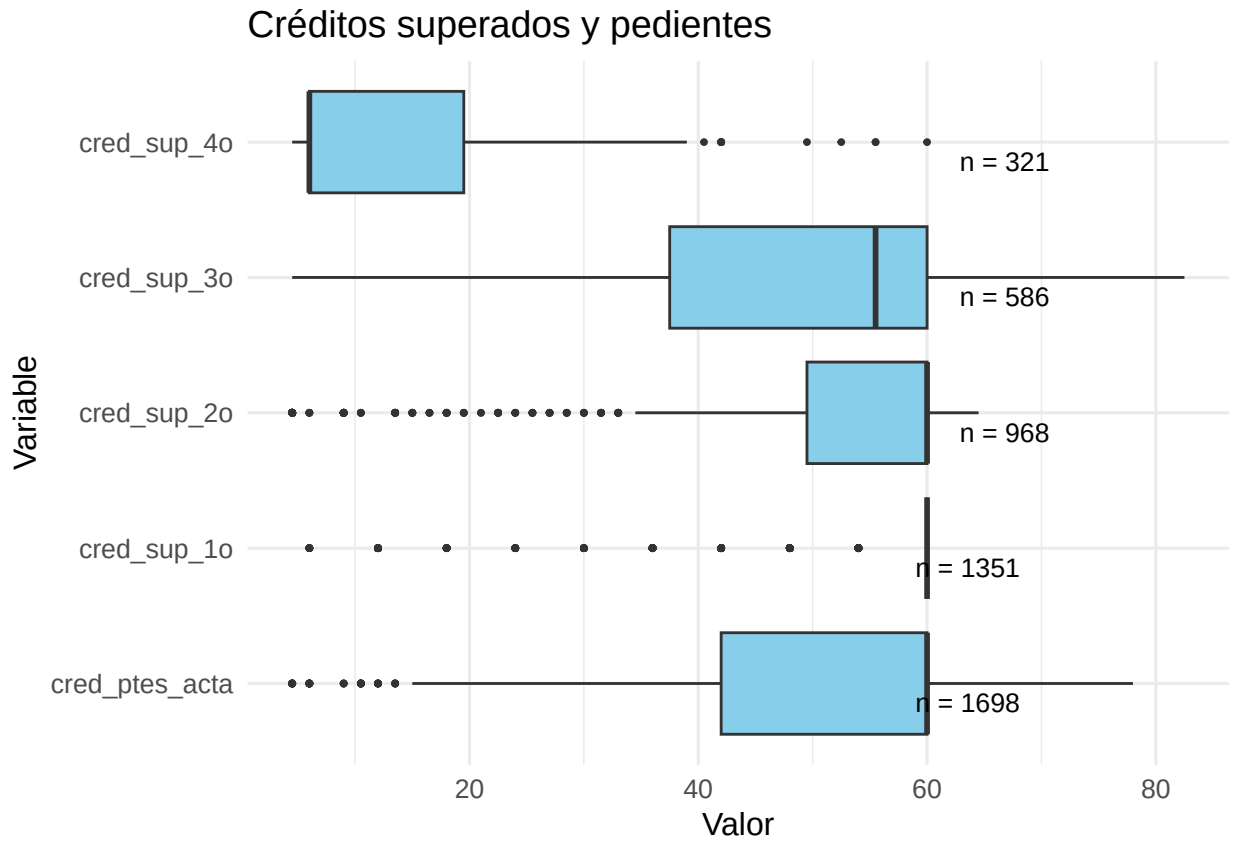


```r
academicas=select(academicas, -cred_pend_sup_tit)
```

```r
df_long=academicas %>%
  select(all_of(grupo_creditos_moderados))

df_long[df_long==0]=NA

df_long=df_long%>%
  pivot_longer(cols = everything(), names_to = "Variable", values_to = "Valor")

conteos=df_long %>%
  filter(!is.na(Valor)) %>%
  group_by(Variable) %>%
  summarise(n = n())


ggplot(df_long, aes(x = Valor, y = fct_reorder(Variable, -Valor, .fun = median, na.rm = TRUE))) +
  geom_boxplot(fill = "skyblue", outlier.size = 0.7) +
  geom_text(data = conteos, aes(x = Inf, y = fct_reorder(Variable, -n), label = paste0("n = ", n)),
            hjust = 3,  size = 3.5, inherit.aes = FALSE,vjust = 1.5) +
  labs(title = "Créditos superados y pedientes", x = "Valor", y = "Variable") +
  theme_minimal(base_size = 12)
```

```
## Warning: `fct_reorder()` removing 3846 missing values.
## i Use `.na_rm = TRUE` to silence this message.
## i Use `.na_rm = FALSE` to preserve NAs.
```

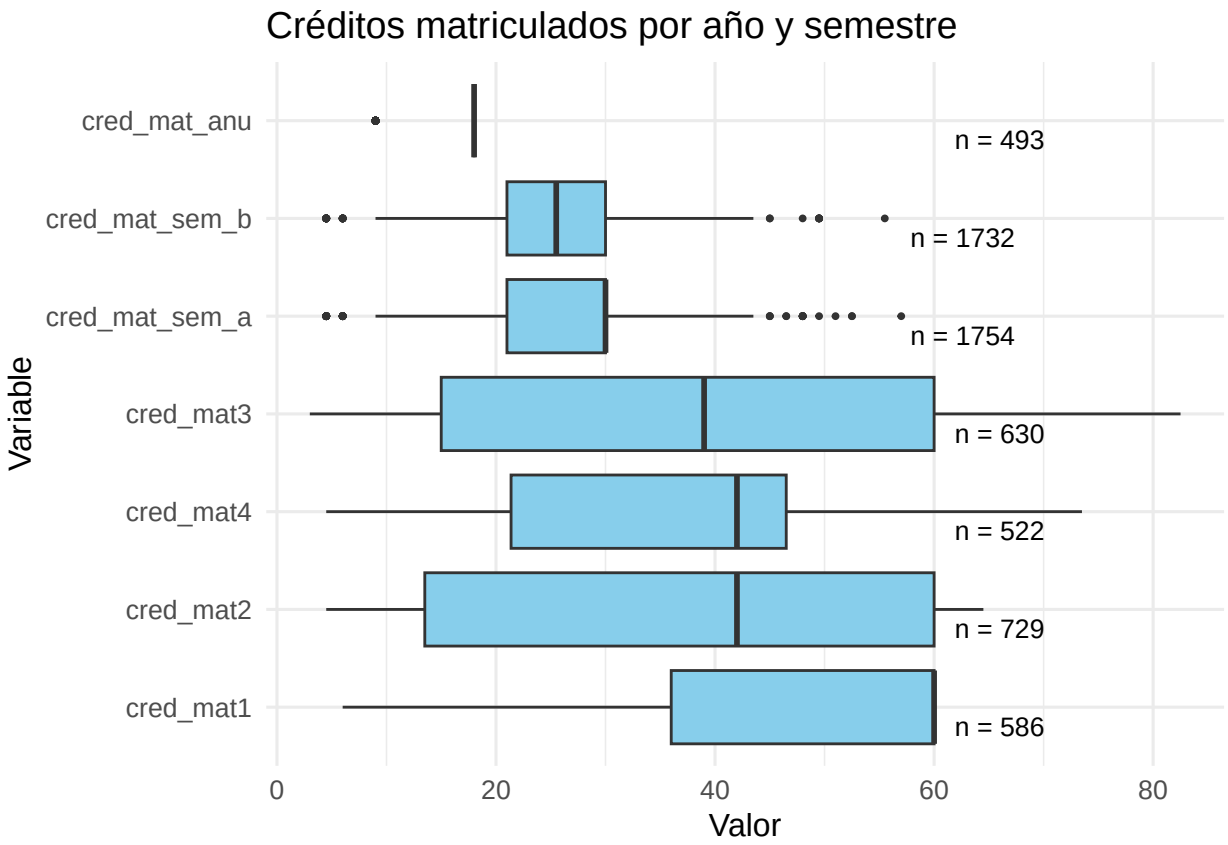
```
## Warning: Removed 3846 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```




```
df_long=academicas %>%
  select(all_of(grupo_creditos_individuales))

df_long[df_long==0]=NA

df_long=df_long%>%
  pivot_longer(cols = everything(), names_to = "Variable", values_to = "Valor")

conteos=df_long %>%
  filter(!is.na(Valor)) %>%
  group_by(Variable) %>%
  summarise(n = n())


ggplot(df_long, aes(x = Valor, y = fct_reorder(Variable, -Valor, .fun = median, na.rm = TRUE))) +
  geom_boxplot(fill = "skyblue", outlier.size = 0.7) +
  geom_text(data = conteos, aes(x = Inf, y = fct_reorder(Variable, -n), label = paste0("n = ", n)),
            hjust = 3,  size = 3.5, inherit.aes = FALSE,vjust = 1.5) +
  labs(title = "Créditos matriculados por año y semestre", x = "Valor", y = "Variable") +
  theme_minimal(base_size = 12)
```

```
## Warning: `fct_reorder()` removing 5832 missing values.
## i Use `.na_rm = TRUE` to silence this message.
## i Use `.na_rm = FALSE` to preserve NAs.

## Warning: Removed 5832 rows containing non-finite outside the scale range
```

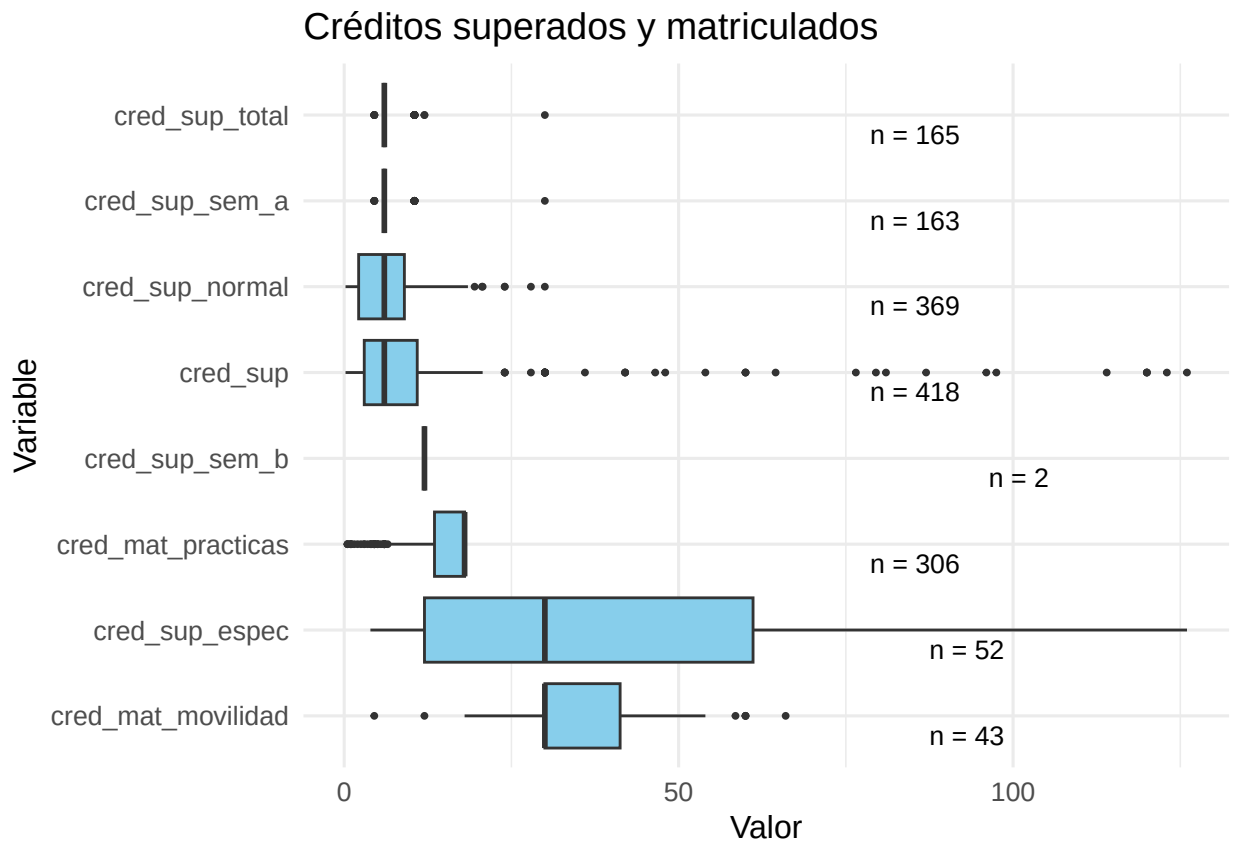
```
## (‘stat_boxplot()‘).
```




```
df_long=academicas %>%
  select(all_of(grupo_creditos_especiales))


df_long[df_long==0]=NA

df_long=df_long%>%
  pivot_longer(cols = everything(), names_to = "Variable", values_to = "Valor")

conteos=df_long %>%
  filter(!is.na(Valor)) %>%
  group_by(Variable) %>%
  summarise(n = n())


ggplot(df_long, aes(x = Valor, y = fct_reorder(Variable, -Valor, .fun = median, na.rm = TRUE))) +
  geom_boxplot(fill = "skyblue", outlier.size = 0.7) +
  geom_text(data = conteos, aes(x = Inf, y = fct_reorder(Variable, -n), label = paste0("n = ", n)),
            hjust = 4,  size = 3.5, inherit.aes = FALSE,vjust = 1.5) +
  labs(title = "Créditos superados y matriculados", x = "Valor", y = "Variable") +
  theme_minimal(base_size = 12)
```

```
## Warning: ‘fct_reorder()‘ removing 12514 missing values.
```

```
## i Use '.na_rm = TRUE' to silence this message.
## i Use '.na_rm = FALSE' to preserve NAs.

## Warning: Removed 12514 rows containing non-finite outside the scale range
## ('stat_boxplot()').
```

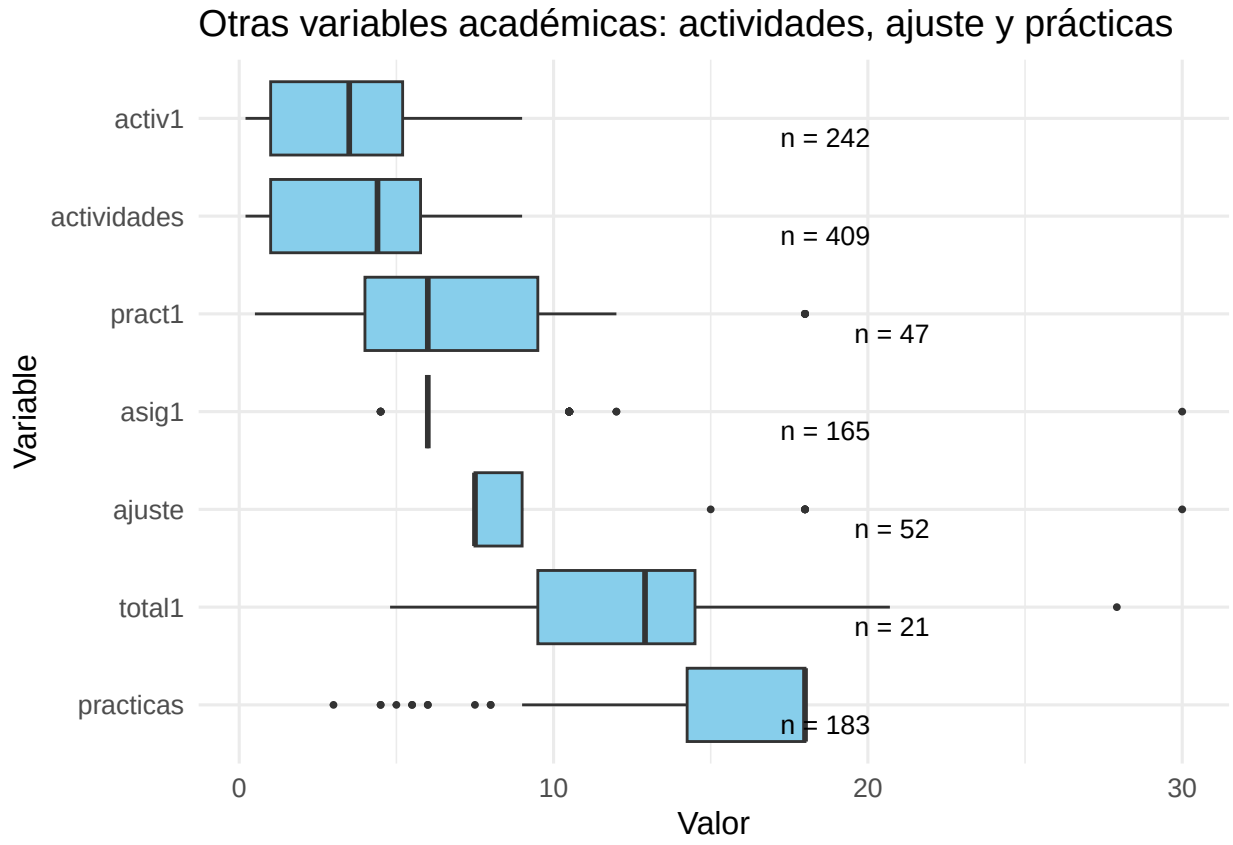



```
df_long=academicas %>%
  select(all_of(grupo_otros_academicos))


df_long[df_long==0]=NA

df_long=df_long%>%
  pivot_longer(cols = everything(), names_to = "Variable", values_to = "Valor")

conteos=df_long %>%
  filter(!is.na(Valor)) %>%
  group_by(Variable) %>%
  summarise(n = n())


ggplot(df_long, aes(x = Valor, y = fct_reorder(Variable, -Valor, .fun = median, na.rm = TRUE))) +
  geom_boxplot(fill = "skyblue", outlier.size = 0.7) +
  geom_text(data = conteos, aes(x = Inf, y = fct_reorder(Variable, -n), label = paste0("n = ", n)),
            hjust = 5,  size = 3.5, inherit.aes = FALSE,vjust = 1.5) +
```

```
  labs(title = "Otras variables académicas: actividades, ajuste y prácticas", x = "Valor", y = "Variable") +
  theme_minimal(base_size = 12)
```

```
## Warning: `fct_reorder()` removing 11159 missing values.
## i Use `.na_rm = TRUE` to silence this message.
## i Use `.na_rm = FALSE` to preserve NAs.
```

```
## Warning: Removed 11159 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```

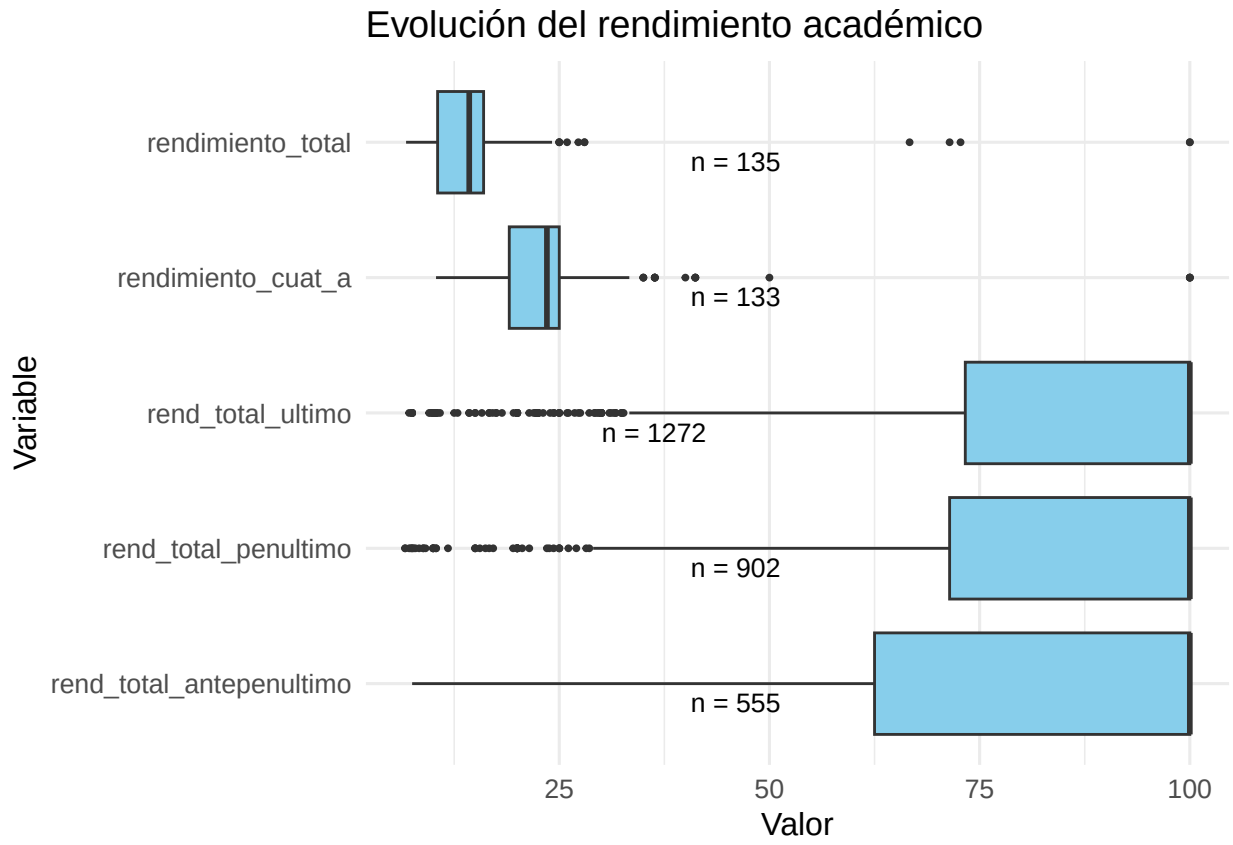



```
df_long=academicas %>%
  select(all_of(rendimientos))

df_long[df_long==0]=NA

df_long=df_long%>%
  pivot_longer(cols = everything(), names_to = "Variable", values_to = "Valor")

conteos=df_long %>%
  filter(!is.na(Valor)) %>%
  group_by(Variable) %>%
  summarise(n = n())
```

```r
ggplot(df_long, aes(x = Valor, y = fct_reorder(Variable, -Valor, .fun = median, na.rm = TRUE))) +
  geom_boxplot(fill = "skyblue", outlier.size = 0.7) +
  geom_text(data = conteos, aes(x = Inf, y = fct_reorder(Variable, -n), label = paste0("n = ", n)),
            hjust = 6,  size = 3.5, inherit.aes = FALSE,vjust = 1.5) +
  labs(title = "Evolución del rendimiento académico", x = "Valor", y = "Variable") +
  theme_minimal(base_size = 12)
```

```
## Warning: `fct_reorder()` removing 5773 missing values.
## i Use `.na_rm = TRUE` to silence this message.
## i Use `.na_rm = FALSE` to preserve NAs.
```

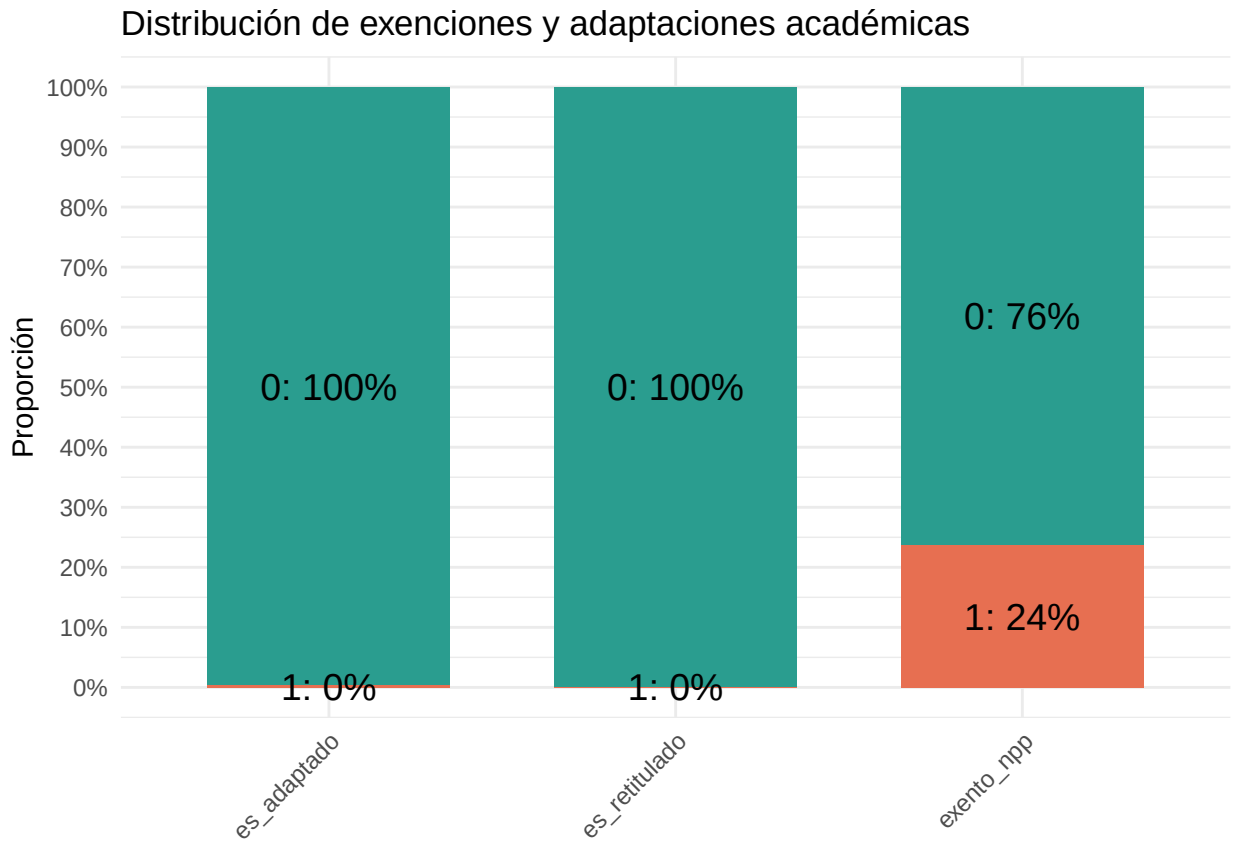
```
## Warning: Removed 5773 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



Binarias de las académicas.

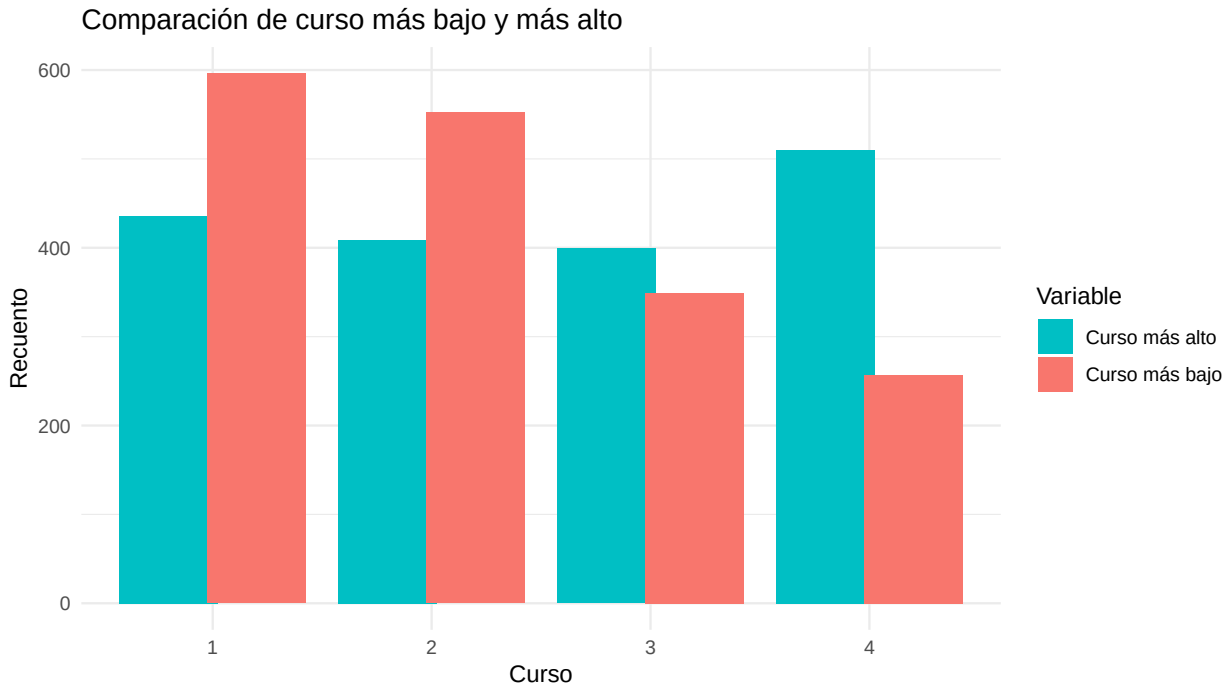
```r
binarias_1=c("exento_npp", "es_adaptado","es_retitulado")

crear_plot_binarias=function(vars) {
  df=academicas %>%
    select(all_of(vars)) %>%
    pivot_longer(cols = everything(), names_to = "Variable", values_to = "Valor") %>%
    filter(!is.na(Valor)) %>%
    group_by(Variable, Valor) %>%
    summarise(n = n(), .groups = "drop") %>%
```

```
    group_by(Variable) %>%
    mutate(Proporcion = n / sum(n),
           Etiqueta = paste0(Valor, ": ", round(Proporcion * 100), "%"))

  ggplot(df, aes(x = Variable, y = Proporcion, fill = as.factor(Valor))) +
    geom_col(position = "stack", width = 0.7) +
    geom_text(aes(label = Etiqueta), position = position_stack(vjust = 0.5),
              color = "black", size = 5) +
    scale_y_continuous(labels = scales::percent_format(accuracy = 1),
                       breaks = seq(0, 1, 0.1), limits = c(0, 1)) +
    scale_fill_manual(values = c("0" = "#2a9d8f", "1" = "#e76f51")) +
    labs(title = "Distribución de exenciones y adaptaciones académicas",
         y = "Proporción", x = NULL) +
    theme_minimal() +
    theme(legend.position = "none",
          axis.text.x = element_text(angle = 45, hjust = 1))
}


# Crear los gráficos
crear_plot_binarias(binarias_1)
```



Comparación de cursos.

```r
df_barras=academicas %>%
  select(curso_mas_alto, curso_mas_bajo) %>%
  pivot_longer(cols = everything(), names_to = "Variable", values_to = "Valor") %>%
  filter(!is.na(Valor)) %>%
  group_by(Variable, Valor) %>%
  summarise(Recuento = n(), .groups = "drop")


df_barras$Valor=factor(df_barras$Valor, levels = 1:4)


ggplot(df_barras, aes(x = Valor, y = Recuento, fill = Variable)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.8)) +
  labs(title = "Comparación de curso más bajo y más alto",
       x = "Curso", y = "Recuento") +
  scale_fill_manual(values = c("#00BFC4", "#F8766D"), labels = c("Curso más alto", "Curso más bajo")) +
  theme_minimal()
```



Digital

```r
summary(poliformat)
```

```
##    dni_hash         pft_events_2024_7_est pft_events_2024_8_est
##  Length:1754        Min.   : 0.000        Min.   :  0.000
##  Class :character   1st Qu.: 0.000        1st Qu.:  0.000
##  Mode  :character   Median : 0.000        Median :  0.000
##                     Mean   : 0.118        Mean   :  1.256
```

```
##                    3rd Qu.: 0.000        3rd Qu.:  0.000
##                    Max.   :25.000        Max.   :198.000
##  pft_events_2024_9_est pft_events_2024_10_est pft_events_2024_11_est
##  Min.   :   0.0        Min.   :   0.0         Min.   :   0.0
##  1st Qu.:   0.0        1st Qu.:   0.0         1st Qu.:   0.0
##  Median :   0.0        Median :   0.0         Median :   0.0
##  Mean   : 105.5        Mean   : 207.4         Mean   : 110.8
##  3rd Qu.: 174.0        3rd Qu.: 394.8         3rd Qu.: 183.8
##  Max.   :1178.0        Max.   :2014.0         Max.   :1442.0
##  pft_events_2024_12_est pft_events_2025_1_est pft_days_logged_2024_7_est
##  Min.   :   0.0         Min.   :   0.0        Min.   :0.00000
##  1st Qu.:   0.0         1st Qu.:   0.0        1st Qu.:0.00000
##  Median :   0.0         Median :   0.0        Median :0.00000
##  Mean   : 245.0         Mean   : 178.6        Mean   :0.02223
##  3rd Qu.: 448.8         3rd Qu.: 318.0        3rd Qu.:0.00000
##  Max.   :2782.0         Max.   :2910.0        Max.   :3.00000
##  pft_days_logged_2024_8_est pft_days_logged_2024_9_est
##  Min.   : 0.0000            Min.   :  0.0
##  1st Qu.: 0.0000            1st Qu.:  0.0
##  Median : 0.0000            Median :  0.0
##  Mean   : 0.1956            Mean   : 14.7
##  3rd Qu.: 0.0000            3rd Qu.: 27.0
##  Max.   :16.0000            Max.   :113.0
##  pft_days_logged_2024_10_est pft_days_logged_2024_11_est
##  Min.   :  0.00              Min.   :  0.00
##  1st Qu.:  0.00              1st Qu.:  0.00
##  Median :  0.00              Median :  0.00
##  Mean   : 24.06              Mean   : 14.78
##  3rd Qu.: 48.00              3rd Qu.: 27.00
##  Max.   :175.00              Max.   :172.00
##  pft_days_logged_2024_12_est pft_days_logged_2025_1_est pft_visits_2024_7_est
##  Min.   :  0.0               Min.   :  0.00             Min.   :0.0000
##  1st Qu.:  0.0               1st Qu.:  0.00             1st Qu.:0.0000
##  Median :  0.0               Median :  0.00             Median :0.0000
##  Mean   : 22.9               Mean   : 15.74             Mean   :0.0325
##  3rd Qu.: 45.0               3rd Qu.: 31.00             3rd Qu.:0.0000
##  Max.   :145.0               Max.   :103.00             Max.   :9.0000
##  pft_visits_2024_8_est pft_visits_2024_9_est pft_visits_2024_10_est
##  Min.   : 0.0000       Min.   :  0.0         Min.   :  0.00
##  1st Qu.: 0.0000       1st Qu.:  0.0         1st Qu.:  0.00
##  Median : 0.0000       Median :  0.0         Median :  0.00
##  Mean   : 0.2799       Mean   : 24.2         Mean   : 45.55
##  3rd Qu.: 0.0000       3rd Qu.: 42.0         3rd Qu.: 88.00
##  Max.   :41.0000       Max.   :264.0         Max.   :559.00
##  pft_visits_2024_11_est pft_visits_2024_12_est pft_visits_2025_1_est
##  Min.   :  0.00         Min.   :  0.00         Min.   :  0.00
##  1st Qu.:  0.00         1st Qu.:  0.00         1st Qu.:  0.00
##  Median :  0.00         Median :  0.00         Median :  0.00
##  Mean   : 26.06         Mean   : 46.56         Mean   : 33.34
##  3rd Qu.: 44.00         3rd Qu.: 86.75         3rd Qu.: 61.00
##  Max.   :487.00         Max.   :461.00         Max.   :431.00
##  pft_total_minutes_2024_7_est pft_total_minutes_2024_8_est
##  Min.   :  0.000              Min.   :   0.0
##  1st Qu.:  0.000              1st Qu.:   0.0
```

```
##  Median :  0.000              Median :   0.0
##  Mean   :  1.207              Mean   :  11.1
##  3rd Qu.:  0.000              3rd Qu.:   0.0
##  Max.   :197.637              Max.   :1115.5
##  pft_total_minutes_2024_9_est pft_total_minutes_2024_10_est
##  Min.   :    0                Min.   :    0
##  1st Qu.:    0                1st Qu.:    0
##  Median :    0                Median :    0
##  Mean   : 1047                Mean   : 1976
##  3rd Qu.: 1780                3rd Qu.: 3657
##  Max.   :11284                Max.   :23055
##  pft_total_minutes_2024_11_est pft_total_minutes_2024_12_est
##  Min.   :    0                 Min.   :    0
##  1st Qu.:    0                 1st Qu.:    0
##  Median :    0                 Median :    0
##  Mean   : 1114                 Mean   : 2076
##  3rd Qu.: 1891                 3rd Qu.: 3806
##  Max.   :18176                 Max.   :21340
##  pft_total_minutes_2025_1_est n_wifi_days_2024_7 n_wifi_days_2024_8
##  Min.   :    0                Min.   : 0.0000    Min.   : 0.00000
##  1st Qu.:    0                1st Qu.: 0.0000    1st Qu.: 0.00000
##  Median :    0                Median : 0.0000    Median : 0.00000
##  Mean   : 1543                Mean   : 0.5707    Mean   : 0.04675
##  3rd Qu.: 2827                3rd Qu.: 0.0000    3rd Qu.: 0.00000
##  Max.   :18271                Max.   :23.0000    Max.   :14.00000
##  n_wifi_days_2024_9 n_wifi_days_2024_10 n_wifi_days_2024_11 n_wifi_days_2024_12
##  Min.   : 0.000     Min.   : 0.00       Min.   : 0.0000     Min.   : 0.000
##  1st Qu.: 0.000     1st Qu.: 0.00       1st Qu.: 0.0000     1st Qu.: 0.000
##  Median : 9.000     Median :13.00       Median : 0.0000     Median : 5.000
##  Mean   : 7.184     Mean   :10.16       Mean   : 0.6608     Mean   : 4.958
##  3rd Qu.:12.000     3rd Qu.:16.00       3rd Qu.: 0.0000     3rd Qu.: 8.000
##  Max.   :29.000     Max.   :28.00       Max.   :22.0000     Max.   :29.000
##  n_wifi_days_2025_1 n_resource_days_2024_7_est n_resource_days_2024_8_est
##  Min.   : 0.000     Min.   :0.000000           Min.   : 0.00000
##  1st Qu.: 0.000     1st Qu.:0.000000           1st Qu.: 0.00000
##  Median : 3.000     Median :0.000000           Median : 0.00000
##  Mean   : 4.359     Mean   :0.007982           Mean   : 0.09863
##  3rd Qu.: 7.000     3rd Qu.:0.000000           3rd Qu.: 0.00000
##  Max.   :27.000     Max.   :2.000000           Max.   :15.00000
##  n_resource_days_2024_9_est n_resource_days_2024_10_est
##  Min.   : 0.00              Min.   :  0.00
##  1st Qu.: 0.00              1st Qu.:  0.00
##  Median : 0.00              Median :  0.00
##  Mean   :11.04              Mean   : 17.79
##  3rd Qu.:21.00              3rd Qu.: 35.00
##  Max.   :80.00              Max.   :113.00
##  n_resource_days_2024_11_est n_resource_days_2024_12_est
##  Min.   :  0.000             Min.   :  0.00
##  1st Qu.:  0.000             1st Qu.:  0.00
##  Median :  0.000             Median :  0.00
##  Mean   :  9.499             Mean   : 15.34
##  3rd Qu.: 17.000             3rd Qu.: 30.00
##  Max.   :101.000             Max.   :103.00
##  n_resource_days_2025_1_est resource_events_2024_7_est
```

```
##  Min.   : 0.00              Min.   : 0.00000
##  1st Qu.: 0.00              1st Qu.: 0.00000
##  Median : 0.00              Median : 0.00000
##  Mean   :10.65              Mean   : 0.03478
##  3rd Qu.:21.00              3rd Qu.: 0.00000
##  Max.   :71.00              Max.   :12.00000
##  resource_events_2024_8_est resource_events_2024_9_est
##  Min.   : 0.0000            Min.   :  0.00
##  1st Qu.: 0.0000            1st Qu.:  0.00
##  Median : 0.0000            Median :  0.00
##  Mean   : 0.3141            Mean   : 31.97
##  3rd Qu.: 0.0000            3rd Qu.: 54.00
##  Max.   :78.0000            Max.   :302.00
##  resource_events_2024_10_est resource_events_2024_11_est
##  Min.   :  0.00              Min.   :  0.00
##  1st Qu.:  0.00              1st Qu.:  0.00
##  Median :  0.00              Median :  0.00
##  Mean   : 68.47              Mean   : 31.38
##  3rd Qu.:126.00              3rd Qu.: 54.00
##  Max.   :617.00              Max.   :370.00
##  resource_events_2024_12_est resource_events_2025_1_est
##  Min.   :  0.0               Min.   :  0.00
##  1st Qu.:  0.0               1st Qu.:  0.00
##  Median :  0.0               Median :  0.00
##  Mean   : 72.1               Mean   : 56.46
##  3rd Qu.:132.0               3rd Qu.:104.00
##  Max.   :676.0               Max.   :570.00
##  pft_assignment_submissions_2024_9_est pft_assignment_submissions_2024_10_est
##  Min.   : 0.0000                       Min.   : 0.0000
##  1st Qu.: 0.0000                       1st Qu.: 0.0000
##  Median : 0.0000                       Median : 0.0000
##  Mean   : 0.1802                       Mean   : 0.9088
##  3rd Qu.: 0.0000                       3rd Qu.: 1.0000
##  Max.   :10.0000                       Max.   :17.0000
##  pft_assignment_submissions_2024_11_est pft_assignment_submissions_2024_12_est
##  Min.   : 0.0000                        Min.   : 0.000
##  1st Qu.: 0.0000                        1st Qu.: 0.000
##  Median : 0.0000                        Median : 0.000
##  Mean   : 0.6688                        Mean   : 1.436
##  3rd Qu.: 0.0000                        3rd Qu.: 2.000
##  Max.   :17.0000                        Max.   :17.000
##  pft_assignment_submissions_2025_1_est pft_test_submissions_2024_9_est
##  Min.   : 0.0000                       Min.   : 0.000
##  1st Qu.: 0.0000                       1st Qu.: 0.000
##  Median : 0.0000                       Median : 0.000
##  Mean   : 0.4544                       Mean   : 1.014
##  3rd Qu.: 0.0000                       3rd Qu.: 0.000
##  Max.   :12.0000                       Max.   :27.000
##  pft_test_submissions_2024_10_est pft_test_submissions_2024_11_est
##  Min.   : 0.000                   Min.   : 0.000
##  1st Qu.: 0.000                   1st Qu.: 0.000
##  Median : 0.000                   Median : 0.000
##  Mean   : 1.993                   Mean   : 1.158
##  3rd Qu.: 2.000                   3rd Qu.: 2.000
```

```
##  Max.   :32.000                   Max.   :32.000
##  pft_test_submissions_2024_12_est pft_test_submissions_2025_1_est
##  Min.   : 0.000                   Min.   : 0.000
##  1st Qu.: 0.000                   1st Qu.: 0.000
##  Median : 0.000                   Median : 0.000
##  Mean   : 2.123                   Mean   : 1.178
##  3rd Qu.: 2.000                   3rd Qu.: 1.000
##  Max.   :56.000                   Max.   :35.000
##     abandono       evento_poliformat_asg_media loggins_asg_media
##  Min.   :0.00000   Min.   :   0.0              Min.   :  0.00
##  1st Qu.:0.00000   1st Qu.:   0.0              1st Qu.:  0.00
##  Median :0.00000   Median :   0.0              Median :  0.00
##  Mean   :0.02965   Mean   : 156.3              Mean   : 16.87
##  3rd Qu.:0.00000   3rd Qu.: 291.2              3rd Qu.: 33.19
##  Max.   :1.00000   Max.   :2245.2              Max.   :114.00
##  visitas_asg_media minutos_asg_media n_res_day_asg_media n_res_events_asg_media
##  Min.   :  0.00    Min.   :    0     Min.   : 0.00       Min.   :  0.00
##  1st Qu.:  0.00    1st Qu.:    0     1st Qu.: 0.00       1st Qu.:  0.00
##  Median :  0.00    Median :    0     Median : 0.00       Median :  0.00
##  Mean   : 32.11    Mean   : 1421     Mean   :11.78       Mean   : 47.93
##  3rd Qu.: 60.00    3rd Qu.: 2672     3rd Qu.:23.66       3rd Qu.: 92.00
##  Max.   :355.25    Max.   :18398     Max.   :85.25       Max.   :753.00
##  asign_asg_media  test_asg_media    minutos_asg
##  Min.   :0.0000   Min.   : 0.000   Min.   :    0
##  1st Qu.:0.0000   1st Qu.: 0.000   1st Qu.:    0
##  Median :0.0000   Median : 0.000   Median :    2
##  Mean   :0.6556   Mean   : 1.397   Mean   : 7769
##  3rd Qu.:0.8333   3rd Qu.: 1.833   3rd Qu.:14514
##  Max.   :9.5000   Max.   :18.000   Max.   :82186
```

```
colnames(poliformat)
```

```
##  [1] "dni_hash"
##  [2] "pft_events_2024_7_est"
##  [3] "pft_events_2024_8_est"
##  [4] "pft_events_2024_9_est"
##  [5] "pft_events_2024_10_est"
##  [6] "pft_events_2024_11_est"
##  [7] "pft_events_2024_12_est"
##  [8] "pft_events_2025_1_est"
##  [9] "pft_days_logged_2024_7_est"
## [10] "pft_days_logged_2024_8_est"
## [11] "pft_days_logged_2024_9_est"
## [12] "pft_days_logged_2024_10_est"
## [13] "pft_days_logged_2024_11_est"
## [14] "pft_days_logged_2024_12_est"
## [15] "pft_days_logged_2025_1_est"
## [16] "pft_visits_2024_7_est"
## [17] "pft_visits_2024_8_est"
## [18] "pft_visits_2024_9_est"
## [19] "pft_visits_2024_10_est"
## [20] "pft_visits_2024_11_est"
## [21] "pft_visits_2024_12_est"
## [22] "pft_visits_2025_1_est"
```

```
## [23] "pft_total_minutes_2024_7_est"
## [24] "pft_total_minutes_2024_8_est"
## [25] "pft_total_minutes_2024_9_est"
## [26] "pft_total_minutes_2024_10_est"
## [27] "pft_total_minutes_2024_11_est"
## [28] "pft_total_minutes_2024_12_est"
## [29] "pft_total_minutes_2025_1_est"
## [30] "n_wifi_days_2024_7"
## [31] "n_wifi_days_2024_8"
## [32] "n_wifi_days_2024_9"
## [33] "n_wifi_days_2024_10"
## [34] "n_wifi_days_2024_11"
## [35] "n_wifi_days_2024_12"
## [36] "n_wifi_days_2025_1"
## [37] "n_resource_days_2024_7_est"
## [38] "n_resource_days_2024_8_est"
## [39] "n_resource_days_2024_9_est"
## [40] "n_resource_days_2024_10_est"
## [41] "n_resource_days_2024_11_est"
## [42] "n_resource_days_2024_12_est"
## [43] "n_resource_days_2025_1_est"
## [44] "resource_events_2024_7_est"
## [45] "resource_events_2024_8_est"
## [46] "resource_events_2024_9_est"
## [47] "resource_events_2024_10_est"
## [48] "resource_events_2024_11_est"
## [49] "resource_events_2024_12_est"
## [50] "resource_events_2025_1_est"
## [51] "pft_assignment_submissions_2024_9_est"
## [52] "pft_assignment_submissions_2024_10_est"
## [53] "pft_assignment_submissions_2024_11_est"
## [54] "pft_assignment_submissions_2024_12_est"
## [55] "pft_assignment_submissions_2025_1_est"
## [56] "pft_test_submissions_2024_9_est"
## [57] "pft_test_submissions_2024_10_est"
## [58] "pft_test_submissions_2024_11_est"
## [59] "pft_test_submissions_2024_12_est"
## [60] "pft_test_submissions_2025_1_est"
## [61] "abandono"
## [62] "evento_poliformat_asg_media"
## [63] "loggins_asg_media"
## [64] "visitas_asg_media"
## [65] "minutos_asg_media"
## [66] "n_res_day_asg_media"
## [67] "n_res_events_asg_media"
## [68] "asign_asg_media"
## [69] "test_asg_media"
## [70] "minutos_asg"
```

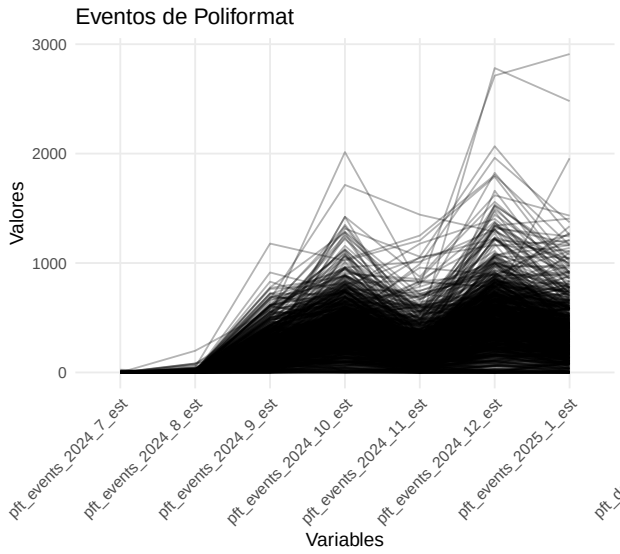
```
#b=skim(poliformat)
#write.csv(b[,c(1:4,10:16)],"skim_poliformat.csv")
```

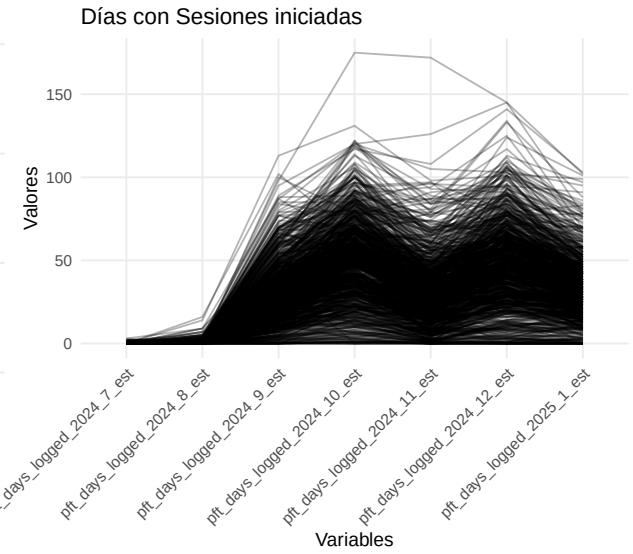
Gráficos de evolución temporal de todas las variables normales.

```r
lineas_multiple=function(df, columnas, titulo) {
  if (is.numeric(columnas)) {
    columnas=names(df)[columnas]
  }

  df_seleccionado=df[, columnas, drop = FALSE]
  df_seleccionado=df_seleccionado[sapply(df_seleccionado, is.numeric)]
  df_seleccionado=unique(df_seleccionado)

  if (ncol(df_seleccionado) == 0) {
    message("No hay columnas numéricas en la selección.")
    return(NULL)
  }

  df_valores=df_seleccionado
  df_valores$ID=1:nrow(df_valores)

  df_melt=reshape2::melt(df_valores, id.vars = "ID",
                         variable.name = "Variable", value.name = "Valor")
  df_melt$Variable=factor(df_melt$Variable, levels = columnas)
  df_melt=unique(df_melt)

  ggplot(df_melt, aes(x = Variable, y = Valor, group = ID)) +
    geom_line(color = "black", alpha = 0.3, size = 0.5) +
    labs(title = titulo, x = "Variables", y = "Valores") +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1),
          panel.grid.minor = element_blank())
}

g1=lineas_multiple(poliformat, c(2:8), "Eventos de Poliformat")
```

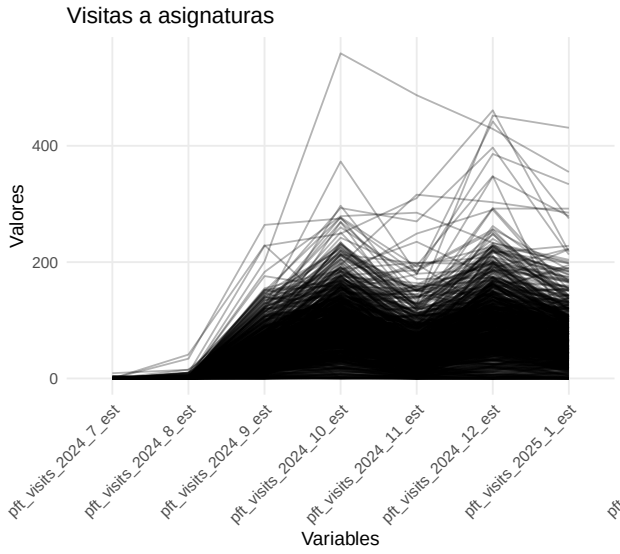
```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

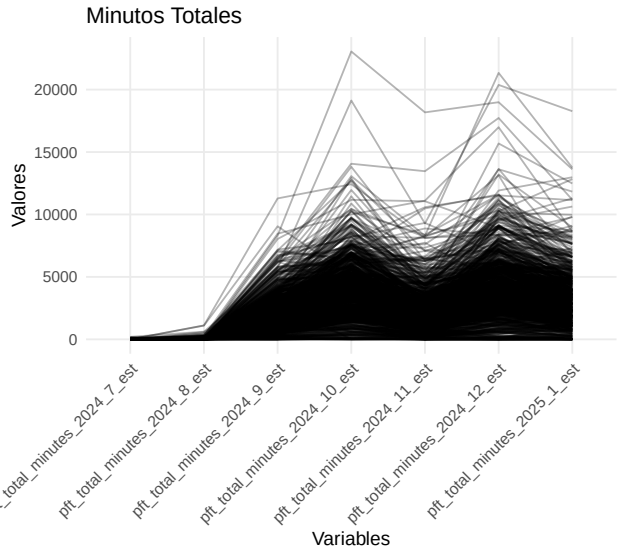
```r
g2=lineas_multiple(poliformat, c(9:15), "Días con Sesiones iniciadas")
g3=lineas_multiple(poliformat, c(16:22), "Visitas a asignaturas")
g4=lineas_multiple(poliformat, c(23:29), "Minutos Totales")
g5=lineas_multiple(poliformat, c(38:43), "Días accedidos a Recursos \nde una asignatura")
g6=lineas_multiple(poliformat, c(44:50), "Acciones en recursos \n de una asignatura")
g7=lineas_multiple(poliformat, c(51:55), "Tareas entregadas en \nuna asignatura")
g8=lineas_multiple(poliformat, c(56:60), "Tests entregados en \nuna asignatura")

grid_1=grid.arrange(g1, g2, ncol = 2)
```



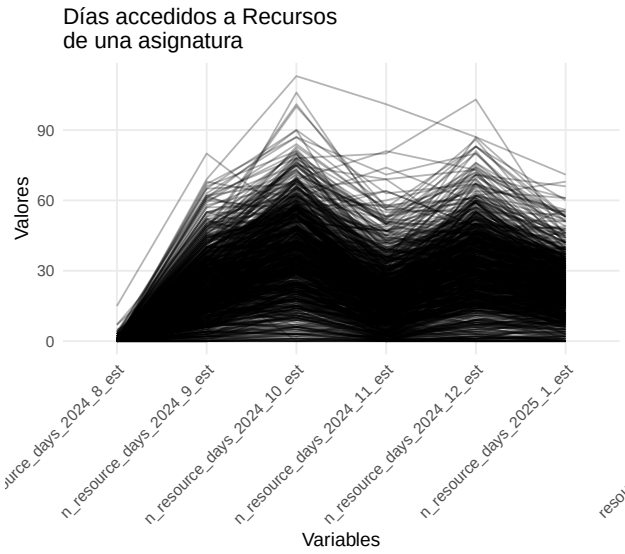


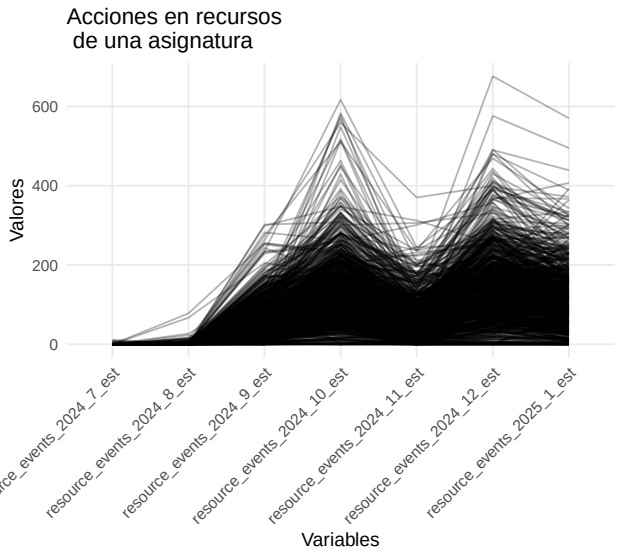

```
grid_2=grid.arrange(g3, g4, ncol = 2)
```



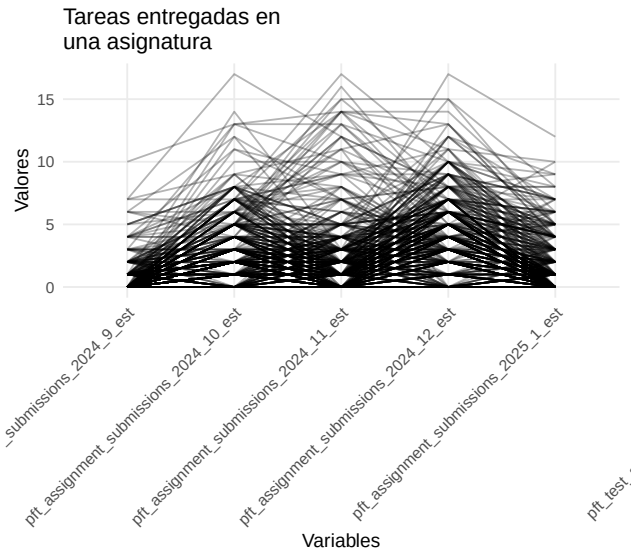


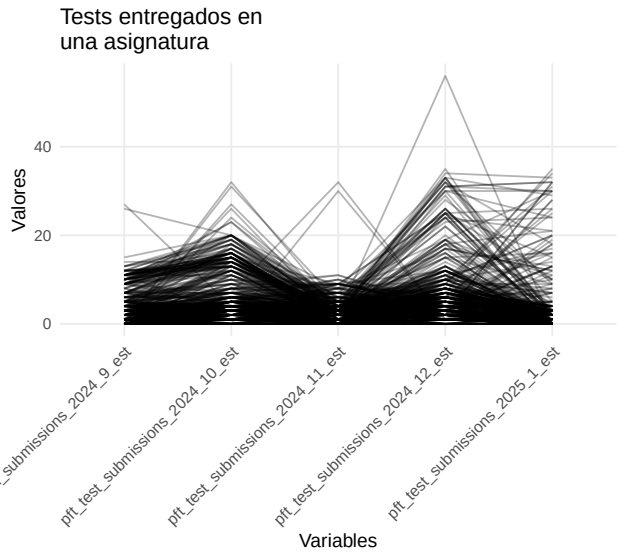

```
grid_3=grid.arrange(g5, g6, ncol = 2)
```






```
grid_4=grid.arrange(g7, g8, ncol = 2)
```

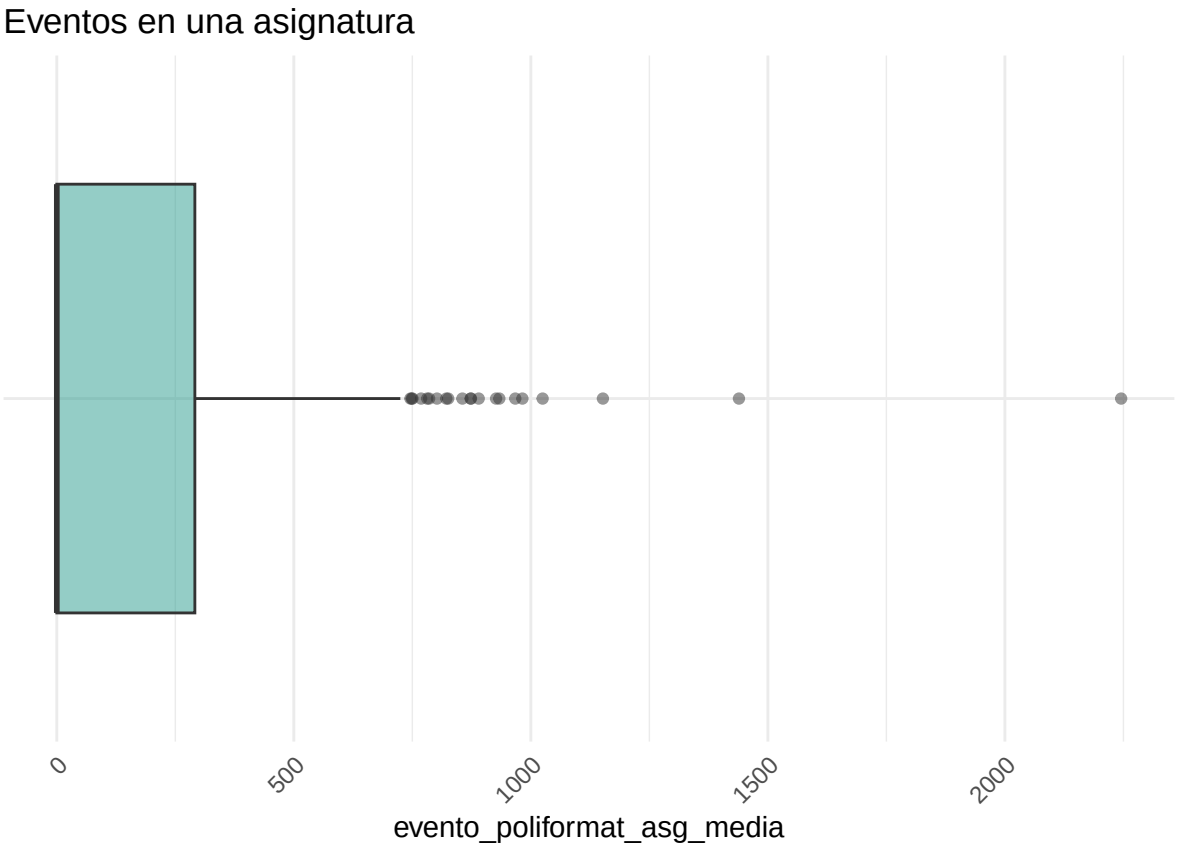




Boxplots de las medias de comportamiento de estudiantes.

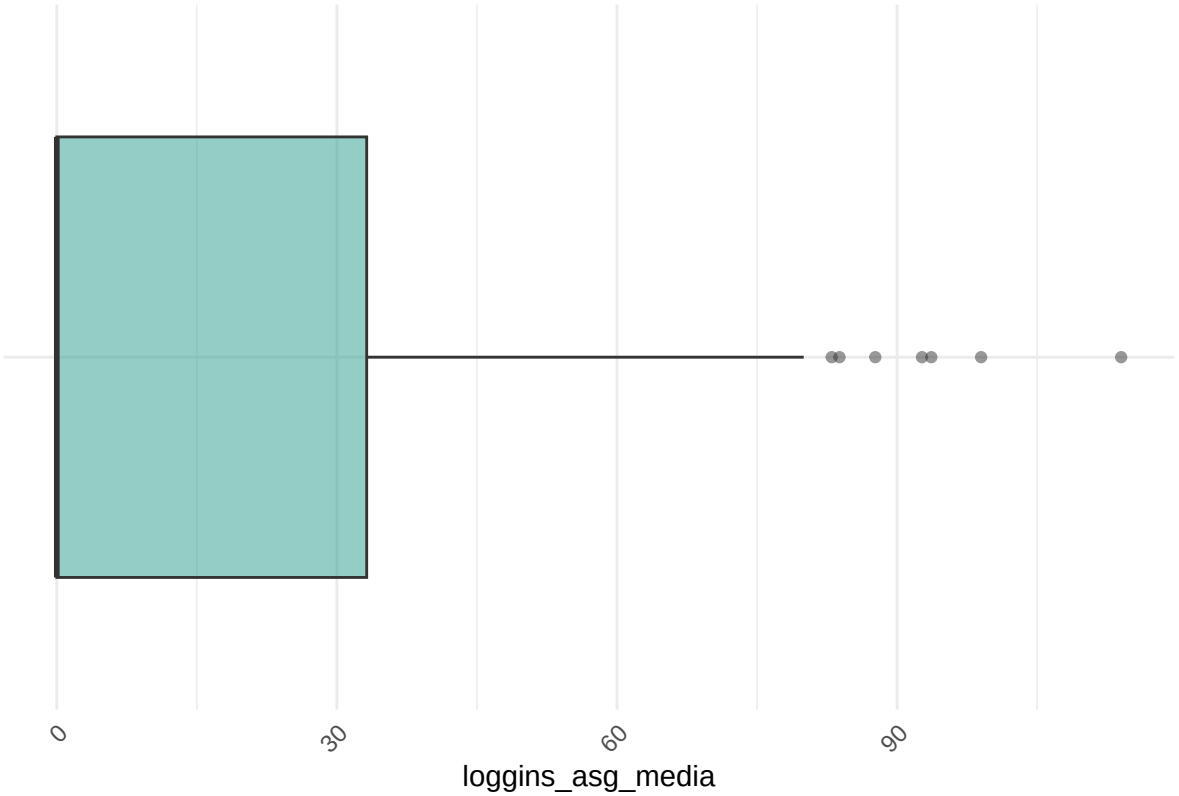
```
variables_asg=function(df, variable, titulo) {

  if (!variable %in% names(df)) {
    stop(paste("La variable", variable, "no existe en el dataframe."))
  }
  if (!is.numeric(df[[variable]])) {
    stop("La variable debe ser numérica.")
  }


  df_filtrado=df[df[[variable]] > 0, ]
```

```
  ggplot(df, aes(x = .data[[variable]], y = "")) +
    geom_boxplot(fill = "#2a9d8f", alpha = 0.5) +
    labs(title = paste0(titulo), x = variable, y = "") +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))

}

variables_asg(poliformat, "evento_poliformat_asg_media", "Eventos en una asignatura")
```

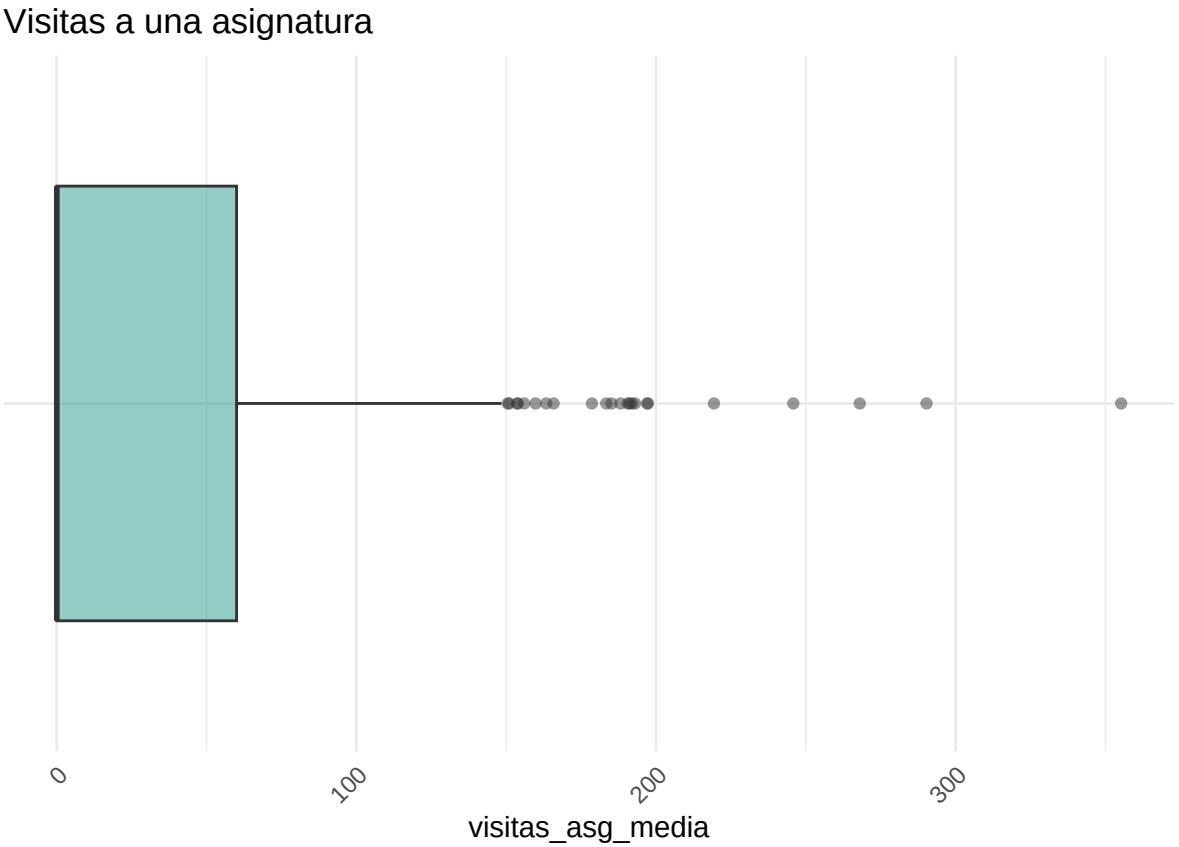



```
variables_asg(poliformat, "loggins_asg_media", "Días con accesos a una asignatura")
```

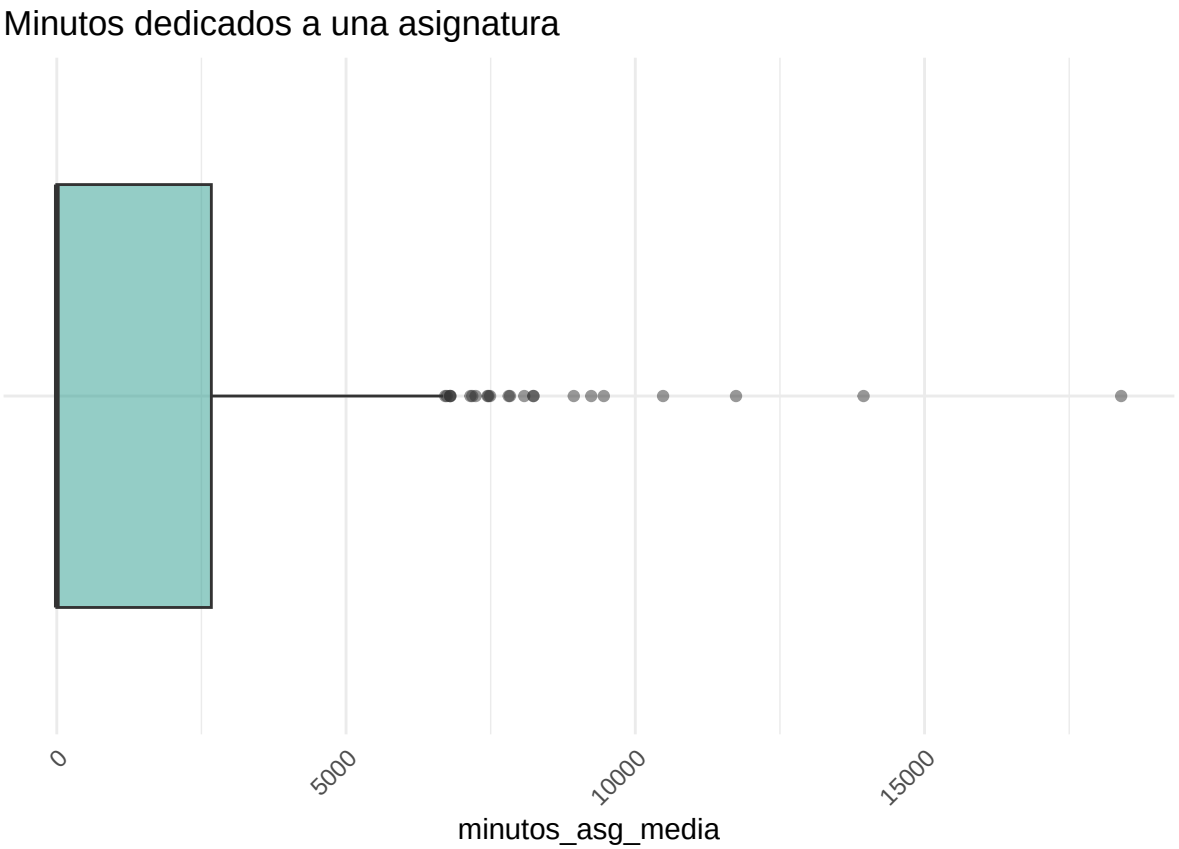
Días con accesos a una asignatura




```
variables_asg(poliformat, "visitas_asg_media", "Visitas a una asignatura")
```

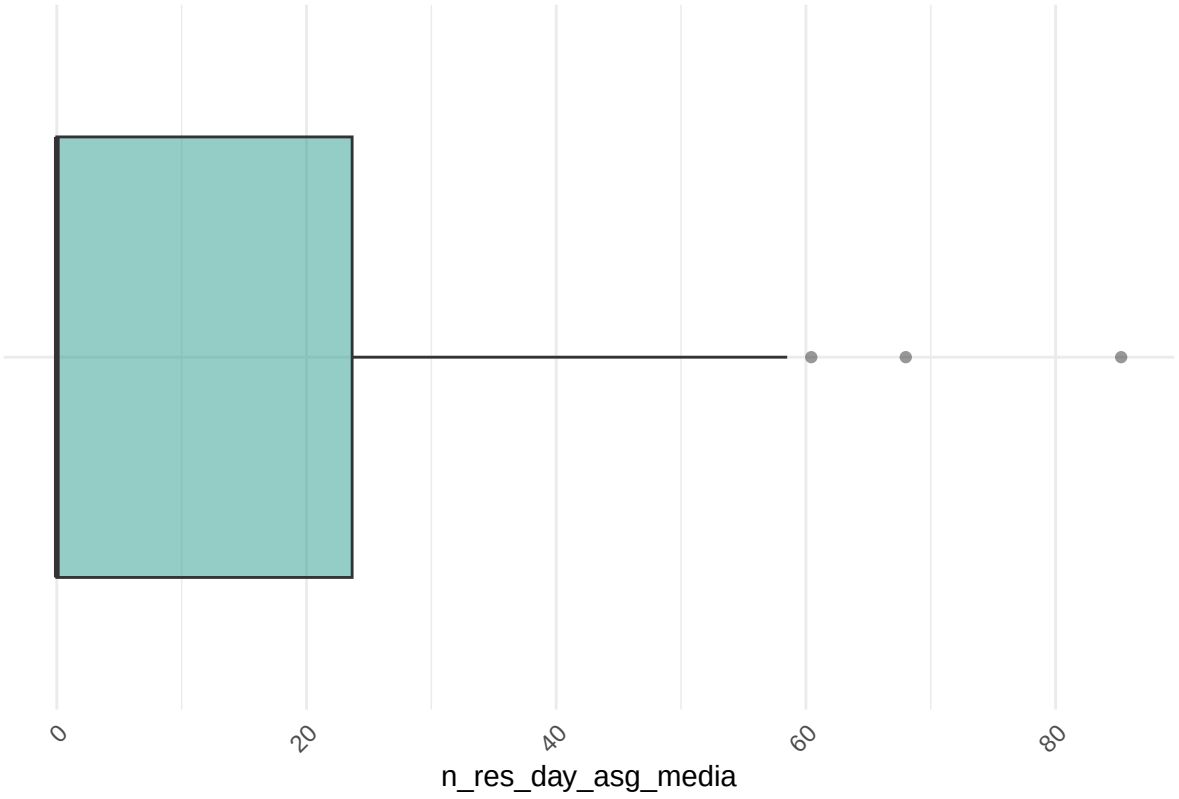



```
variables_asg(poliformat, "minutos_asg_media", "Minutos dedicados a una asignatura")
```

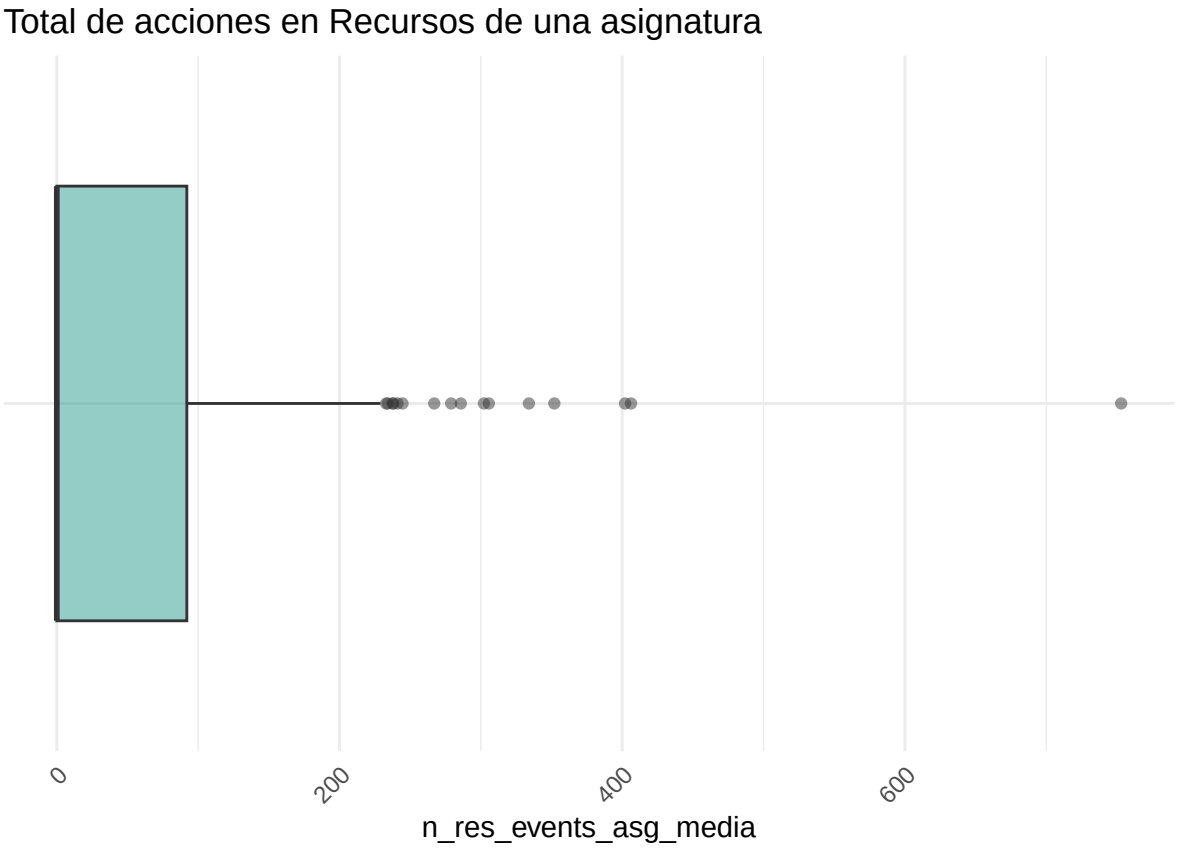



```
variables_asg(poliformat, "n_res_day_asg_media", "Días accedidos a recursos de una asignatura")
```

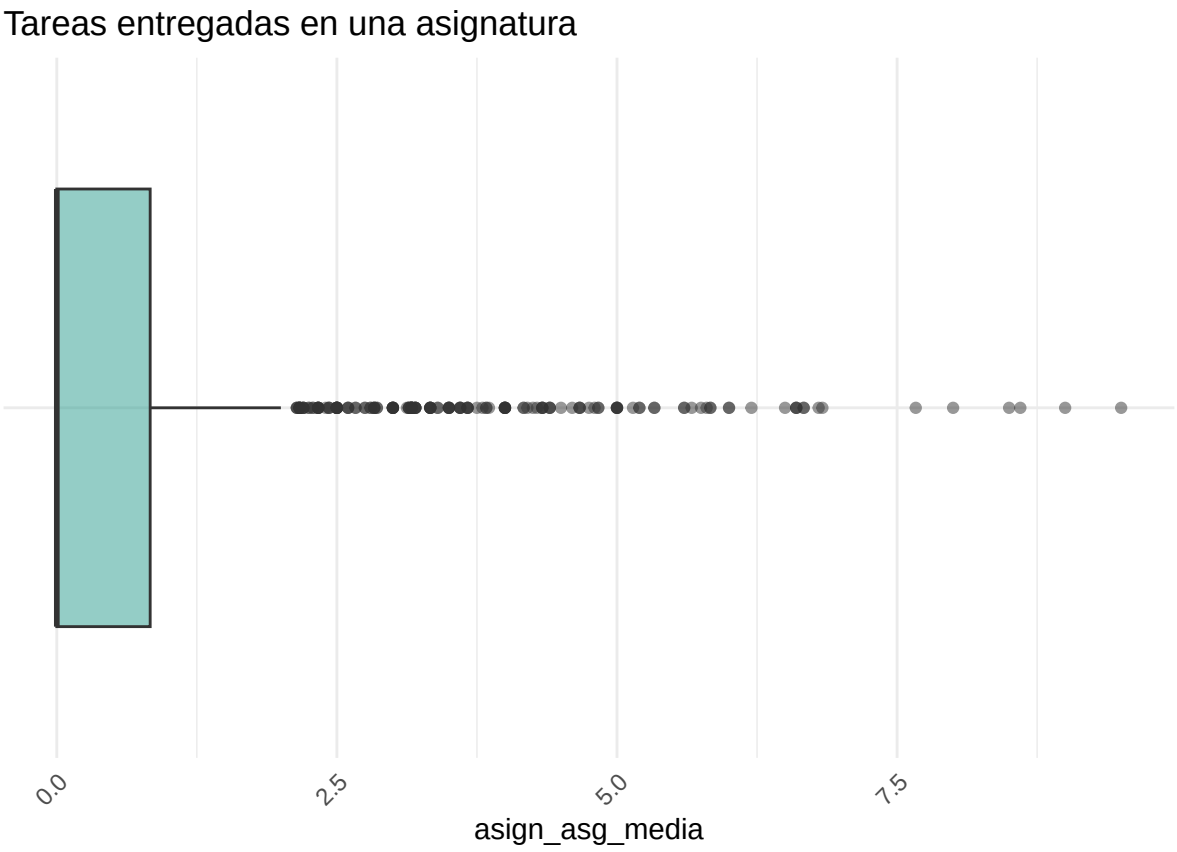
Días accedidos a recursos de una asignatura



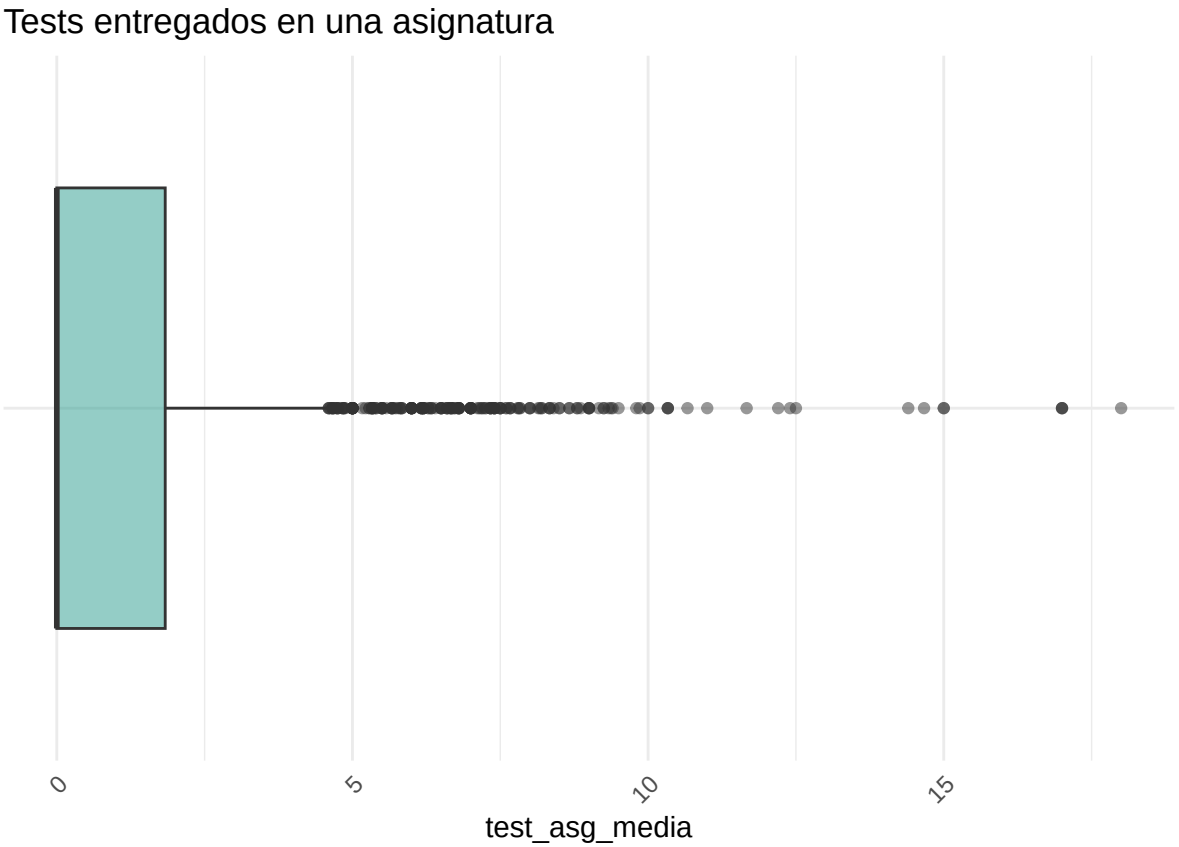

```
variables_asg(poliformat, "n_res_events_asg_media", "Total de acciones en Recursos de una asignatura")
```




```
variables_asg(poliformat, "asign_asg_media", "Tareas entregadas en una asignatura")
```




```
variables_asg(poliformat, "test_asg_media", "Tests entregados en una asignatura")
```

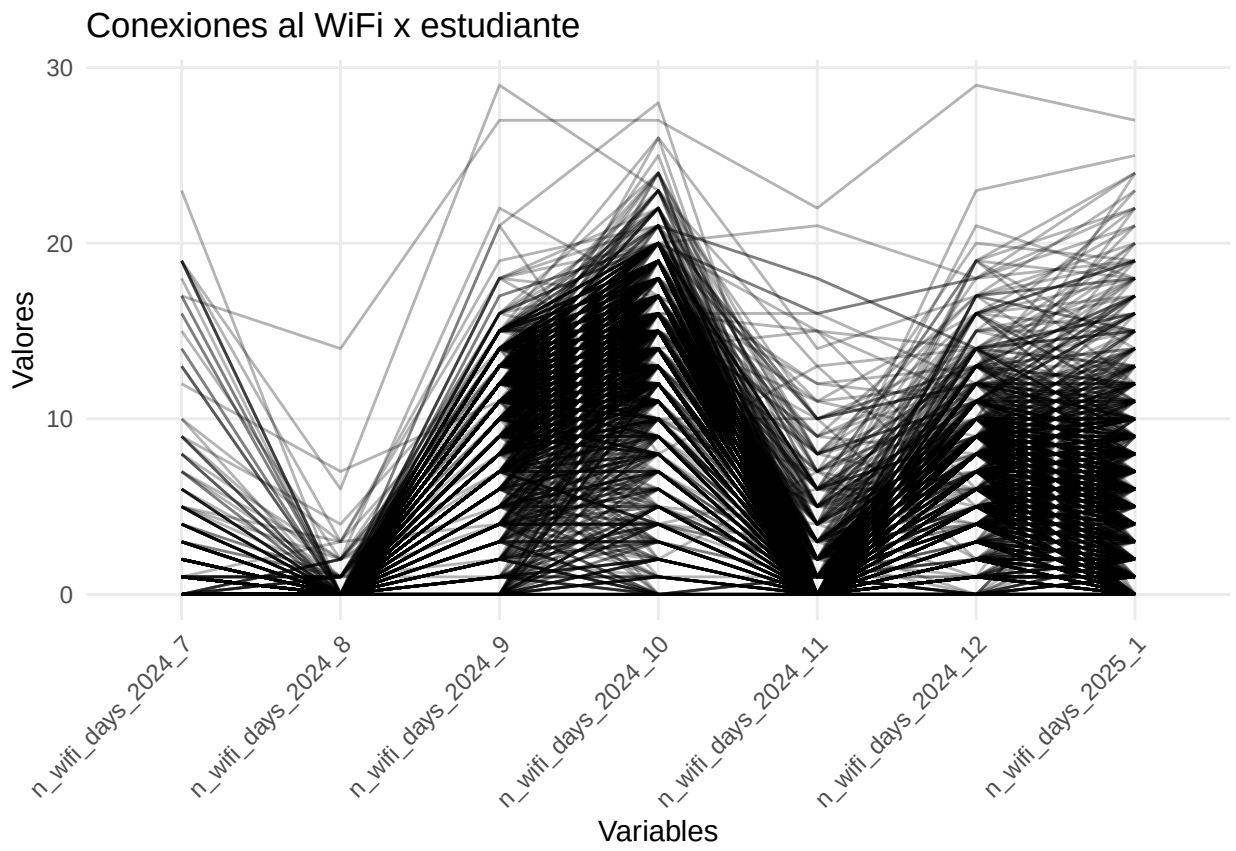



```
colnames(poliformat)
```

```
##  [1] "dni_hash"
##  [2] "pft_events_2024_7_est"
##  [3] "pft_events_2024_8_est"
##  [4] "pft_events_2024_9_est"
##  [5] "pft_events_2024_10_est"
##  [6] "pft_events_2024_11_est"
##  [7] "pft_events_2024_12_est"
##  [8] "pft_events_2025_1_est"
##  [9] "pft_days_logged_2024_7_est"
## [10] "pft_days_logged_2024_8_est"
## [11] "pft_days_logged_2024_9_est"
## [12] "pft_days_logged_2024_10_est"
## [13] "pft_days_logged_2024_11_est"
## [14] "pft_days_logged_2024_12_est"
## [15] "pft_days_logged_2025_1_est"
## [16] "pft_visits_2024_7_est"
## [17] "pft_visits_2024_8_est"
## [18] "pft_visits_2024_9_est"
## [19] "pft_visits_2024_10_est"
## [20] "pft_visits_2024_11_est"
## [21] "pft_visits_2024_12_est"
## [22] "pft_visits_2025_1_est"
## [23] "pft_total_minutes_2024_7_est"
```

```
## [24] "pft_total_minutes_2024_8_est"
## [25] "pft_total_minutes_2024_9_est"
## [26] "pft_total_minutes_2024_10_est"
## [27] "pft_total_minutes_2024_11_est"
## [28] "pft_total_minutes_2024_12_est"
## [29] "pft_total_minutes_2025_1_est"
## [30] "n_wifi_days_2024_7"
## [31] "n_wifi_days_2024_8"
## [32] "n_wifi_days_2024_9"
## [33] "n_wifi_days_2024_10"
## [34] "n_wifi_days_2024_11"
## [35] "n_wifi_days_2024_12"
## [36] "n_wifi_days_2025_1"
## [37] "n_resource_days_2024_7_est"
## [38] "n_resource_days_2024_8_est"
## [39] "n_resource_days_2024_9_est"
## [40] "n_resource_days_2024_10_est"
## [41] "n_resource_days_2024_11_est"
## [42] "n_resource_days_2024_12_est"
## [43] "n_resource_days_2025_1_est"
## [44] "resource_events_2024_7_est"
## [45] "resource_events_2024_8_est"
## [46] "resource_events_2024_9_est"
## [47] "resource_events_2024_10_est"
## [48] "resource_events_2024_11_est"
## [49] "resource_events_2024_12_est"
## [50] "resource_events_2025_1_est"
## [51] "pft_assignment_submissions_2024_9_est"
## [52] "pft_assignment_submissions_2024_10_est"
## [53] "pft_assignment_submissions_2024_11_est"
## [54] "pft_assignment_submissions_2024_12_est"
## [55] "pft_assignment_submissions_2025_1_est"
## [56] "pft_test_submissions_2024_9_est"
## [57] "pft_test_submissions_2024_10_est"
## [58] "pft_test_submissions_2024_11_est"
## [59] "pft_test_submissions_2024_12_est"
## [60] "pft_test_submissions_2025_1_est"
## [61] "abandono"
## [62] "evento_poliformat_asg_media"
## [63] "loggins_asg_media"
## [64] "visitas_asg_media"
## [65] "minutos_asg_media"
## [66] "n_res_day_asg_media"
## [67] "n_res_events_asg_media"
## [68] "asign_asg_media"
## [69] "test_asg_media"
## [70] "minutos_asg"
```

```
lineas_multiple(poliformat, c(30:36), "Conexiones al WiFi x estudiante")
```



Abandono como tal

```
summary(abandono)
```

```
##    dni_hash            asi_left        abandono
##  Length:52          Min.   : 6.00   Min.   :1
##  Class :character   1st Qu.:10.00   1st Qu.:1
##  Mode  :character   Median :10.00   Median :1
##                     Mean   :10.04   Mean   :1
##                     3rd Qu.:10.00   3rd Qu.:1
##                     Max.   :29.00   Max.   :1
##    baja_fecha                              mes
##  Min.   :2024-07-19 00:00:00.00   enero     : 2
##  1st Qu.:2024-07-23 00:00:00.00   julio     :23
##  Median :2024-09-04 00:00:00.00   septiembre:20
##  Mean   :2024-09-01 11:10:23.07   octubre   : 5
##  3rd Qu.:2024-09-17 06:00:00.00   diciembre : 2
##  Max.   :2025-01-28 00:00:00.00
```

```
c=skim(abandono)
```

```r
table(academicas$abandono)
```

```
##
##    0    1
## 1702   52
```

Correlación de los diferentes datasets.

```r
completo=sociodemografia %>%
  left_join(select(academicas,-c("abandono")), by = "dni_hash") %>%
  left_join((select(poliformat,-c("abandono"))), by = "dni_hash")
```

```r
calcular_correlaciones=function(df, bloque = "Dataset") {
  num_vars=names(df)[sapply(df, is.numeric)]
  cat_vars=names(df)[sapply(df, function(x) is.factor(x) || is.character(x))]

  resultados=data.frame(var1 = character(), var2 = character(), tipo = character(), correlacion = numer

  # Numéricas <> Numéricas (Pearson)
  for (i in seq_along(num_vars)) {
    for (j in seq_along(num_vars)) {
      if (i < j) {
        r=cor(df[[num_vars[i]]], df[[num_vars[j]]], use = "pairwise.complete.obs")
        resultados=rbind(resultados, data.frame(
          var1 = num_vars[i], var2 = num_vars[j],
          tipo = "Pearson", correlacion = r
        ))
      }
    }
  }

  # Categóricas <> Categóricas (Cramér's V)
  for (i in seq_along(cat_vars)) {
    for (j in seq_along(cat_vars)) {
      if (i < j) {

        tabla=table(df[[cat_vars[i]]], df[[cat_vars[j]]])
        if (min(dim(tabla)) > 1) {
          cramers=suppressWarnings(assocstats(tabla)$cramer)
          resultados=rbind(resultados, data.frame(
            var1 = cat_vars[i], var2 = cat_vars[j],
            tipo = "Cramér's V", correlacion = cramers
          ))
        }
      }
    }
  }

  # Categóricas <> Numéricas (eta² = proporción de varianza explicada)
  for (cat in cat_vars) {
```

```r
    for (num in num_vars) {
      f=try(summary(aov(df[[num]] ~ as.factor(df[[cat]]))), silent = TRUE)
      if (!inherits(f, "try-error")) {
        ss_total=sum(f[[1]][["Sum Sq"]])
        ss_model=f[[1]][["Sum Sq"]][1]
        eta2=ss_model / ss_total
        resultados=rbind(resultados, data.frame(
          var1 = cat, var2 = num,
          tipo = "Eta²", correlacion = eta2
        ))
      }
    }
  }

  resultados=resultados %>%
    mutate(bloque = bloque) %>%
    arrange(desc(abs(correlacion)))

  return(resultados)
}

cor_aca = calcular_correlaciones(academicas[2:43], "Academicas")
```

```
## Warning in cor(df[[num_vars[i]]], df[[num_vars[j]]], use =
## "pairwise.complete.obs"): La desviación estándar es cero
## Warning in cor(df[[num_vars[i]]], df[[num_vars[j]]], use =
## "pairwise.complete.obs"): La desviación estándar es cero
## Warning in cor(df[[num_vars[i]]], df[[num_vars[j]]], use =
## "pairwise.complete.obs"): La desviación estándar es cero
## Warning in cor(df[[num_vars[i]]], df[[num_vars[j]]], use =
## "pairwise.complete.obs"): La desviación estándar es cero
## Warning in cor(df[[num_vars[i]]], df[[num_vars[j]]], use =
## "pairwise.complete.obs"): La desviación estándar es cero
## Warning in cor(df[[num_vars[i]]], df[[num_vars[j]]], use =
## "pairwise.complete.obs"): La desviación estándar es cero
## Warning in cor(df[[num_vars[i]]], df[[num_vars[j]]], use =
## "pairwise.complete.obs"): La desviación estándar es cero
## Warning in cor(df[[num_vars[i]]], df[[num_vars[j]]], use =
## "pairwise.complete.obs"): La desviación estándar es cero
## Warning in cor(df[[num_vars[i]]], df[[num_vars[j]]], use =
## "pairwise.complete.obs"): La desviación estándar es cero
## Warning in cor(df[[num_vars[i]]], df[[num_vars[j]]], use =
## "pairwise.complete.obs"): La desviación estándar es cero
## Warning in cor(df[[num_vars[i]]], df[[num_vars[j]]], use =
## "pairwise.complete.obs"): La desviación estándar es cero
## Warning in cor(df[[num_vars[i]]], df[[num_vars[j]]], use =
## "pairwise.complete.obs"): La desviación estándar es cero
```

```r
cor_socio = calcular_correlaciones(sociodemografia[2:17], "Sociodemografía")

socio_aca=sociodemografia %>%
  left_join(select(academicas,-c("abandono")), by = "dni_hash")
```

```
cor_socio_aca =calcular_correlaciones(socio_aca, "Completo")
```

```
## Warning in cor(df[[num_vars[i]]], df[[num_vars[j]]], use =
## "pairwise.complete.obs"): La desviación estándar es cero
## Warning in cor(df[[num_vars[i]]], df[[num_vars[j]]], use =
## "pairwise.complete.obs"): La desviación estándar es cero
## Warning in cor(df[[num_vars[i]]], df[[num_vars[j]]], use =
## "pairwise.complete.obs"): La desviación estándar es cero
## Warning in cor(df[[num_vars[i]]], df[[num_vars[j]]], use =
## "pairwise.complete.obs"): La desviación estándar es cero
## Warning in cor(df[[num_vars[i]]], df[[num_vars[j]]], use =
## "pairwise.complete.obs"): La desviación estándar es cero
## Warning in cor(df[[num_vars[i]]], df[[num_vars[j]]], use =
## "pairwise.complete.obs"): La desviación estándar es cero
## Warning in cor(df[[num_vars[i]]], df[[num_vars[j]]], use =
## "pairwise.complete.obs"): La desviación estándar es cero
## Warning in cor(df[[num_vars[i]]], df[[num_vars[j]]], use =
## "pairwise.complete.obs"): La desviación estándar es cero
## Warning in cor(df[[num_vars[i]]], df[[num_vars[j]]], use =
## "pairwise.complete.obs"): La desviación estándar es cero
```

```
fuera=c("dni_hash", "asinom", "baja_fecha")
```

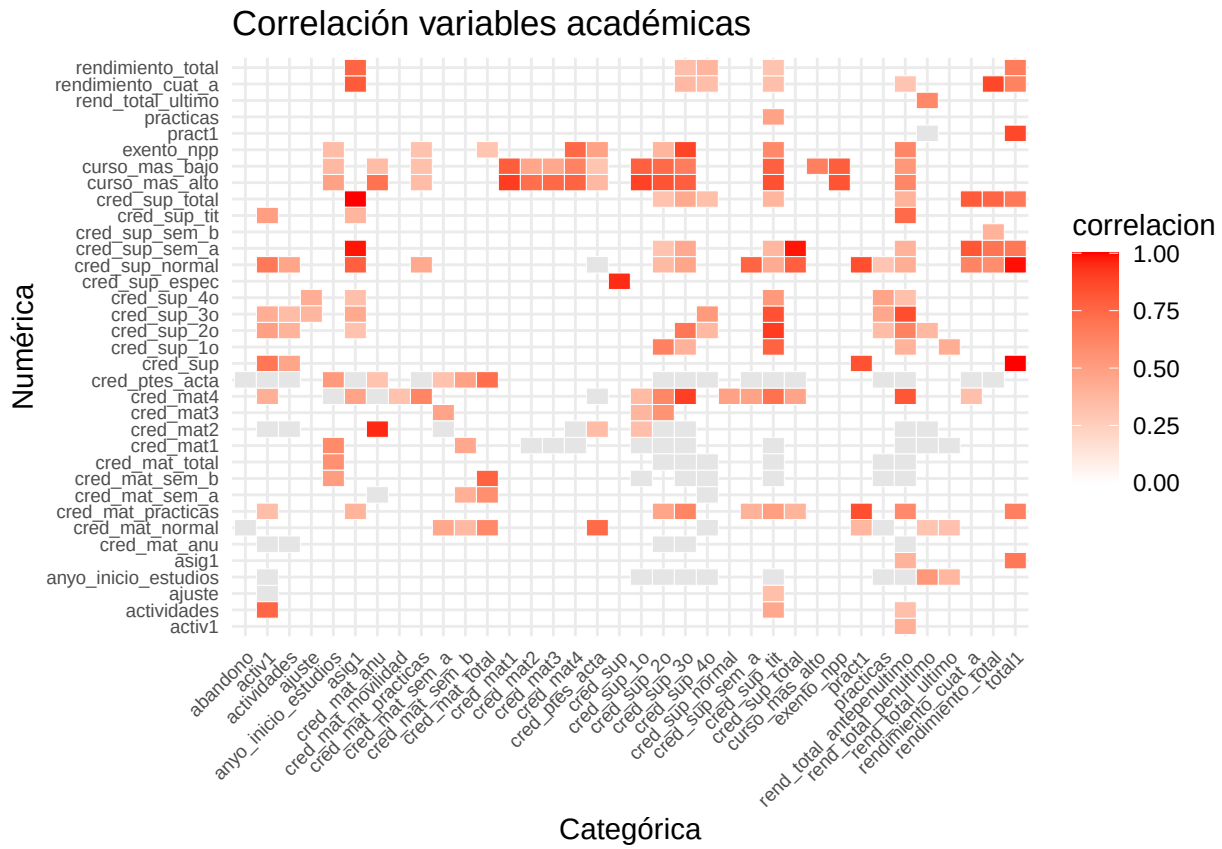
Por el tiempo computacional, no es posible mezclar poliformat con los otros dos datasets.

```
cor_socio_filt=cor_socio %>% filter(abs(correlacion) > 0.3)
cor_aca_filt=cor_aca %>% filter(abs(correlacion) > 0.3)
cor_socio_aca_filt=cor_socio_aca %>% filter(abs(correlacion) > 0.3)

cor_aca_filt=cor_aca_filt[!apply(cor_aca_filt, 1, function(fila) any(fila %in% fuera)), ]
cor_socio_filt=cor_socio_filt[!apply(cor_socio_filt, 1, function(fila) any(fila %in% fuera)), ]
cor_socio_aca_filt=cor_socio_aca_filt[!apply(cor_socio_aca_filt, 1, function(fila) any(fila %in% fuera)
```

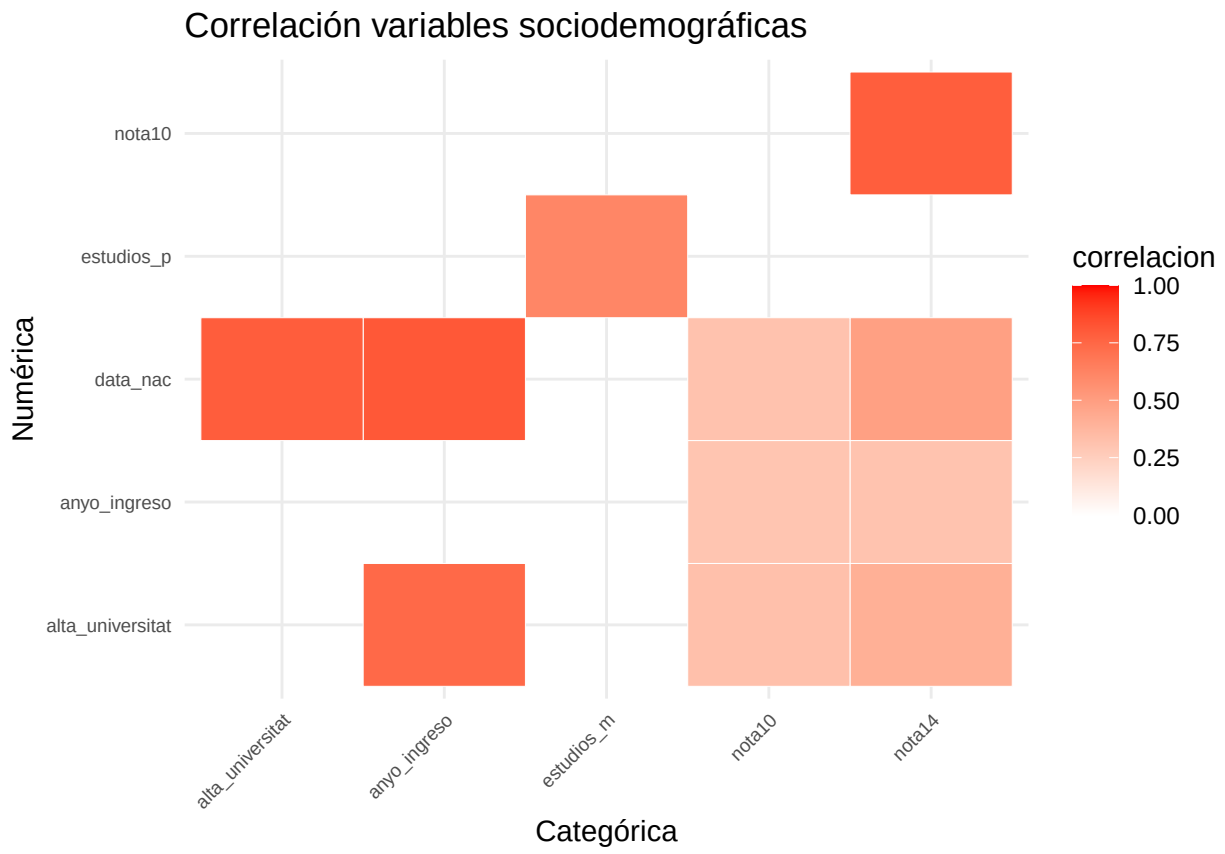
```
melt_academicas=cor_aca_filt %>%
  rename(Numérica = var1, Categórica = var2)


melt_socio_filt=cor_socio_filt %>%
  rename(Numérica = var1, Categórica = var2)

melt_socio_aca_filt=cor_socio_aca_filt %>%
  rename(Numérica = var1, Categórica = var2)
```

Matrices de correlación.

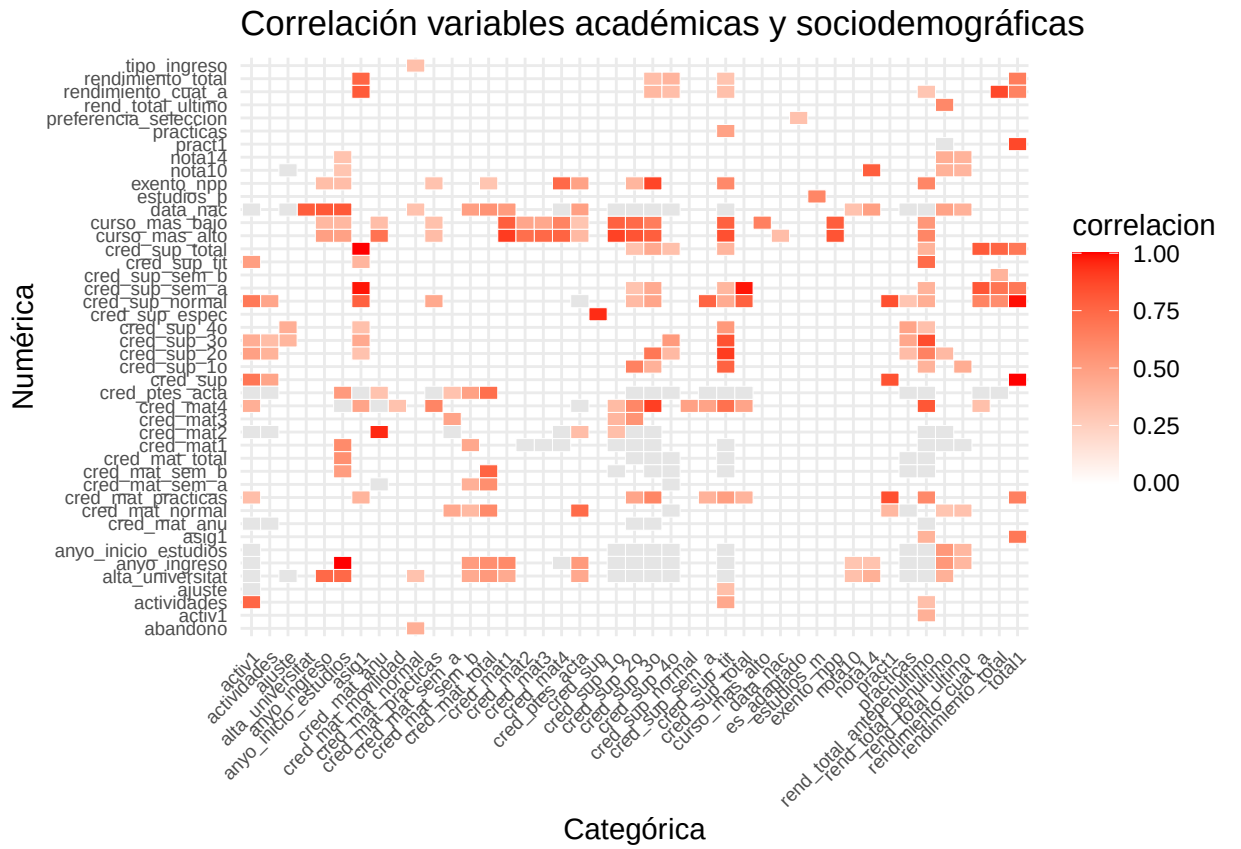
```
ggplot(melt_academicas, aes(x = Categórica, y = Numérica, fill = correlacion)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "white", high = "red", na.value = "grey90", limits = c(0,1)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1,size=7), axis.text.y =element_text(size=7)) +
  labs(title = "Correlación variables académicas")
```




```
ggplot(melt_socio_filt, aes(x = Categórica, y = Numérica, fill = correlacion)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "white", high = "red", na.value = "grey90", limits = c(0,1)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1,size=7), axis.text.y =element_text(size=7)) +
  labs(title = "Correlación variables sociodemográficas")
```




```
ggplot(melt_socio_aca_filt, aes(x = Categórica, y = Numérica, fill = correlacion)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "white", high = "red", na.value = "grey90", limits = c(0,1)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1,size=7), axis.text.y =element_text(size=7)) +
  labs(title = "Correlación variables académicas y sociodemográficas")
```



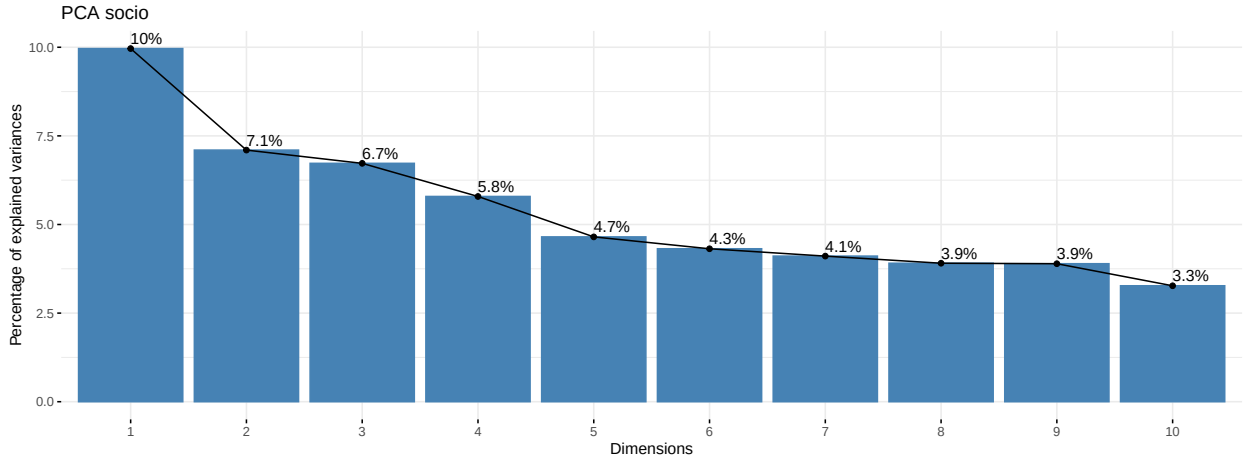
PCA y scree plots de todos los datasets anteriores.

```
num_vars=sociodemografia %>% select(where(is.numeric))
cat_vars=sociodemografia %>% select(where(is.factor) | where(is.character)) %>%
  select(-dni_hash)


cat_dummies=dummy_cols(cat_vars, remove_first_dummy = TRUE, remove_selected_columns = TRUE)


datos_pca=bind_cols(num_vars, cat_dummies)


datos_pca=datos_pca[, sapply(datos_pca, function(x) var(x, na.rm = TRUE) > 0)]


preprocess=preProcess(datos_pca, method = "medianImpute")
datos_imputados=predict(preprocess, newdata = datos_pca)


pca_resultado=prcomp(datos_imputados, scale. = TRUE)


summary(pca_resultado)
```

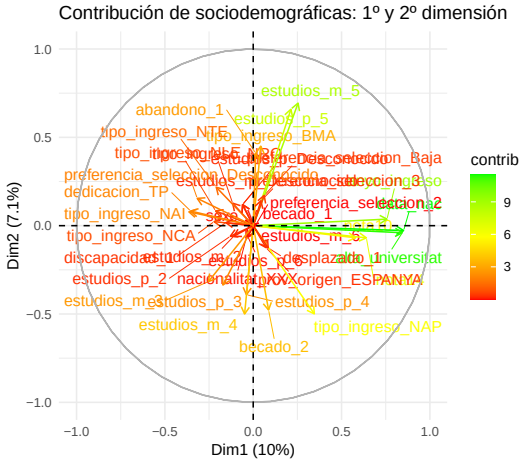
```
## Importance of components:
```

```
##                            PC1     PC2     PC3     PC4    PC5     PC6     PC7
## Standard deviation     1.91998 1.62067 1.57741 1.46376 1.3117 1.26355 1.23285
## Proportion of Variance 0.09963 0.07099 0.06725 0.05791 0.0465 0.04315 0.04108
## Cumulative Proportion  0.09963 0.17062 0.23787 0.29578 0.3423 0.38543 0.42651
##                            PC8     PC9    PC10   PC11    PC12    PC13    PC14
## Standard deviation     1.20230 1.20019 1.10013 1.0830 1.05463 1.05074 1.03688
## Proportion of Variance 0.03907 0.03893 0.03271 0.0317 0.03006 0.02984 0.02906
## Cumulative Proportion  0.46557 0.50450 0.53722 0.5689 0.59897 0.62881 0.65787
##                           PC15    PC16    PC17    PC18   PC19    PC20    PC21
## Standard deviation     1.03443 1.02256 1.02068 0.98617 0.9733 0.96060 0.93164
## Proportion of Variance 0.02892 0.02826 0.02816 0.02628 0.0256 0.02494 0.02346
## Cumulative Proportion  0.68679 0.71505 0.74321 0.76949 0.7951 0.82003 0.84349
##                           PC22    PC23    PC24    PC25    PC26    PC27    PC28
## Standard deviation     0.92366 0.87349 0.85976 0.83867 0.82335 0.72460 0.71791
## Proportion of Variance 0.02306 0.02062 0.01998 0.01901 0.01832 0.01419 0.01393
## Cumulative Proportion  0.86655 0.88717 0.90715 0.92616 0.94448 0.95867 0.97260
##                           PC29    PC30    PC31    PC32    PC33    PC34    PC35
## Standard deviation     0.57238 0.48777 0.43780 0.39240 0.31090 0.07467 0.02020
## Proportion of Variance 0.00885 0.00643 0.00518 0.00416 0.00261 0.00015 0.00001
## Cumulative Proportion  0.98145 0.98788 0.99306 0.99723 0.99984 0.99999 1.00000
##                           PC36      PC37
## Standard deviation     2.101e-15 1.388e-16
## Proportion of Variance 0.000e+00 0.000e+00
## Cumulative Proportion  1.000e+00 1.000e+00
```

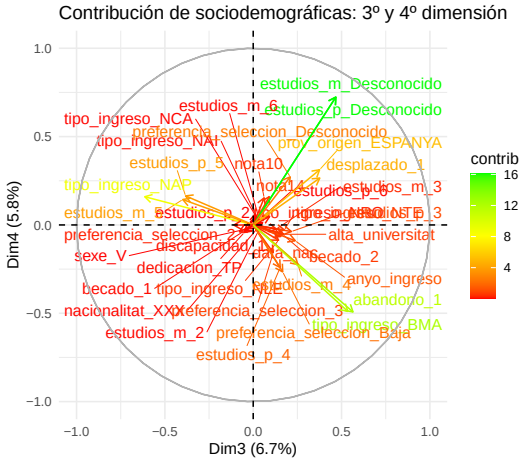
```r
fviz_eig(pca_resultado, addlabels = TRUE, main = "PCA socio")
```




```r
fviz_pca_var(pca_resultado,
             col.var = "contrib",
             gradient.cols = c("red", "yellow", "green"),
             repel = TRUE) +
  labs(title = "Contribución de sociodemográficas: 1º y 2º dimensión") +
  theme_minimal()
```




```
fviz_pca_var(pca_resultado,
             axes=c(3,4),
             col.var = "contrib",
             gradient.cols = c("red", "yellow", "green"),
             repel = TRUE) +
  labs(title = "Contribución de sociodemográficas: 3º y 4º dimensión") +
  theme_minimal()
```

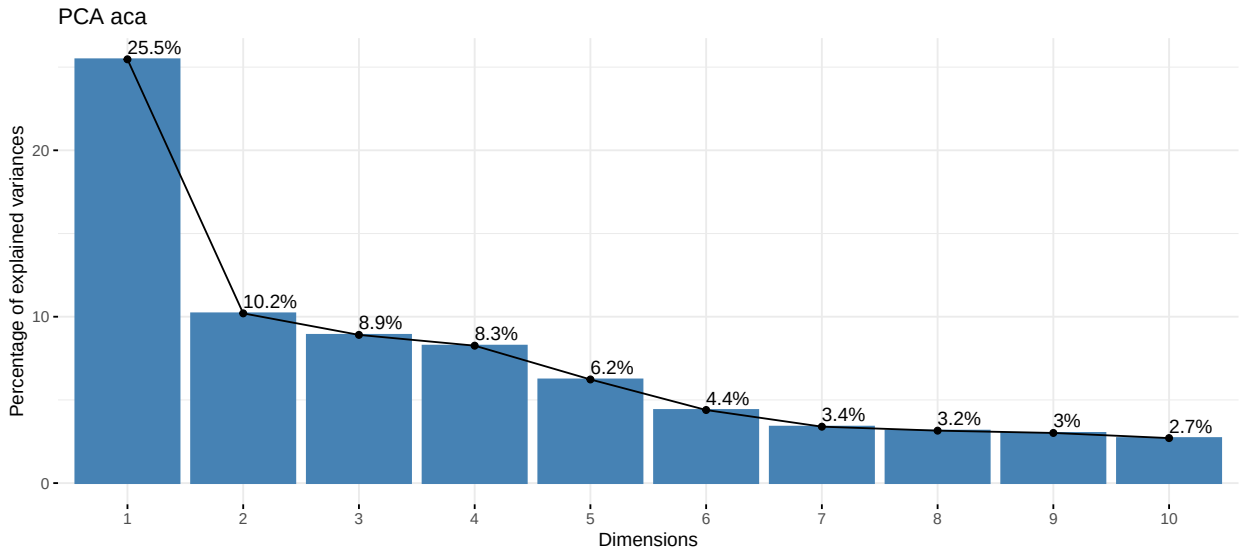



```
num_vars=academicas %>% select(where(is.numeric))
cat_vars=academicas %>% select(where(is.factor) | where(is.character)) %>%
  select(-dni_hash)


cat_dummies=dummy_cols(cat_vars, remove_first_dummy = TRUE, remove_selected_columns = TRUE)


datos_pca=bind_cols(num_vars, cat_dummies)


datos_pca=datos_pca[, sapply(datos_pca, function(x) var(x, na.rm = TRUE) > 0)]
```

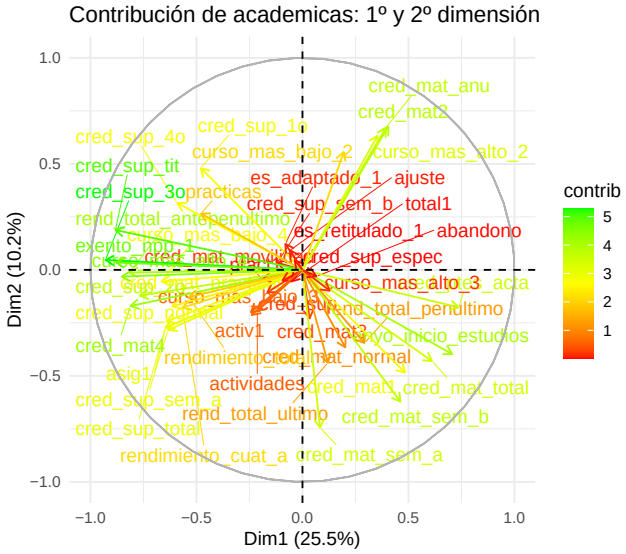
```r
preprocess=preProcess(datos_pca, method = "medianImpute")
datos_imputados=predict(preprocess, newdata = datos_pca)


pca_resultado=prcomp(datos_imputados, scale. = TRUE)


summary(pca_resultado)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     3.4228 2.1663 2.02417 1.94910 1.69266 1.42149 1.24898
## Proportion of Variance 0.2547 0.1020 0.08907 0.08259 0.06228 0.04393 0.03391
## Cumulative Proportion  0.2547 0.3567 0.44577 0.52836 0.59064 0.63457 0.66848
##                            PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     1.20417 1.17751 1.11544 1.04721 1.02901 1.00155 0.97179
## Proportion of Variance 0.03152 0.03014 0.02705 0.02384 0.02302 0.02181 0.02053
## Cumulative Proportion  0.70000 0.73014 0.75719 0.78103 0.80405 0.82586 0.84639
##                           PC15    PC16    PC17    PC18    PC19   PC20    PC21
## Standard deviation     0.93211 0.92436 0.88343 0.82067 0.73795 0.7082 0.65556
## Proportion of Variance 0.01889 0.01857 0.01697 0.01464 0.01184 0.0109 0.00934
## Cumulative Proportion  0.86527 0.88385 0.90082 0.91546 0.92730 0.9382 0.94754
##                           PC22    PC23    PC24    PC25    PC26    PC27    PC28
## Standard deviation     0.62475 0.58873 0.54223 0.52249 0.47619 0.43120 0.40198
## Proportion of Variance 0.00849 0.00753 0.00639 0.00593 0.00493 0.00404 0.00351
## Cumulative Proportion  0.95603 0.96356 0.96995 0.97589 0.98082 0.98486 0.98837
##                           PC29   PC30    PC31    PC32   PC33    PC34    PC35
## Standard deviation     0.35529 0.3392 0.24544 0.23634 0.2145 0.19312 0.17825
## Proportion of Variance 0.00274 0.0025 0.00131 0.00121 0.0010 0.00081 0.00069
## Cumulative Proportion  0.99112 0.9936 0.99493 0.99614 0.9971 0.99795 0.99864
##                           PC36    PC37    PC38    PC39    PC40    PC41
## Standard deviation     0.15673 0.13053 0.09588 0.08541 0.06231 0.02105
## Proportion of Variance 0.00053 0.00037 0.00020 0.00016 0.00008 0.00001
## Cumulative Proportion  0.99918 0.99955 0.99975 0.99991 0.99999 1.00000
##                             PC42      PC43      PC44      PC45     PC46
## Standard deviation     0.0003813 9.404e-15 6.695e-15 4.476e-15 1.52e-15
## Proportion of Variance 0.0000000 0.000e+00 0.000e+00 0.000e+00 0.00e+00
## Cumulative Proportion  1.0000000 1.000e+00 1.000e+00 1.000e+00 1.00e+00
```

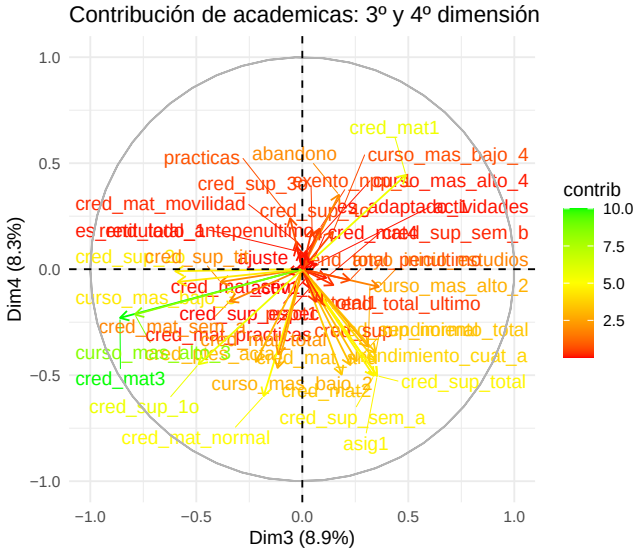
```r
fviz_eig(pca_resultado, addlabels = TRUE, main="PCA aca")
```




```
fviz_pca_var(pca_resultado,
             col.var = "contrib",
             gradient.cols = c("red", "yellow", "green"),
             repel = TRUE) +
  labs(title = "Contribución de academicas: 1º y 2º dimensión") +
  theme_minimal()
```




```
fviz_pca_var(pca_resultado,
             axes=c(3,4),
             col.var = "contrib",
             gradient.cols = c("red", "yellow", "green"),
             repel = TRUE) +
  labs(title = "Contribución de academicas: 3º y 4º dimensión") +
  theme_minimal()
```




```
num_vars=poliformat %>% select(where(is.numeric))
```

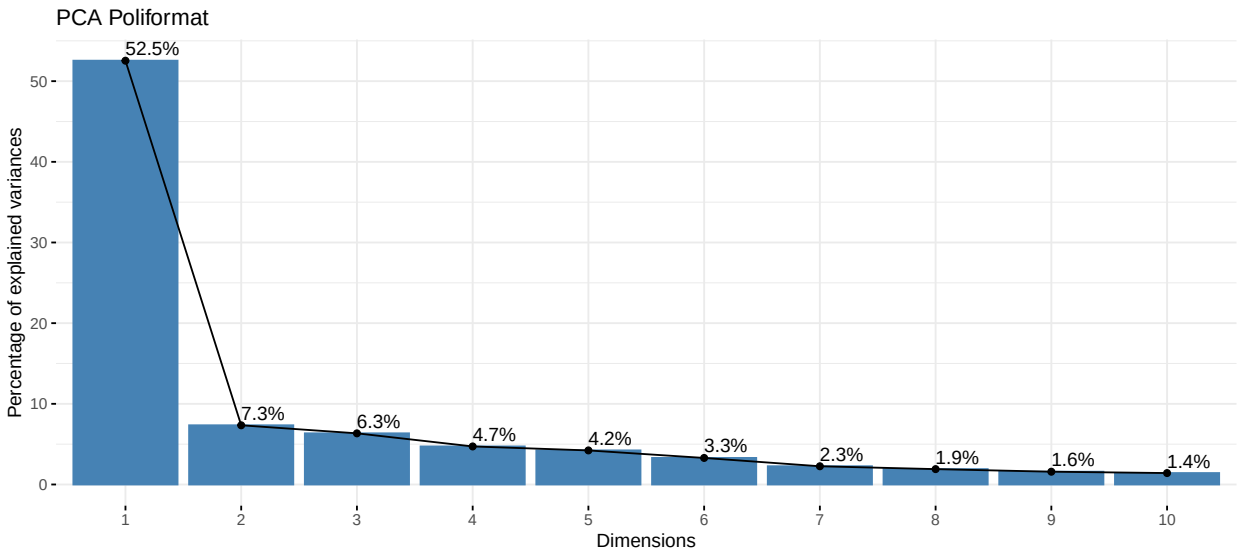
```
## Adding missing grouping variables: `dni_hash`
```

```
num_vars= num_vars[,2:length(num_vars)]



datos_pca=num_vars

preprocess=preProcess(datos_pca, method = "medianImpute")
datos_imputados=predict(preprocess, newdata = datos_pca)


pca_poliformat=prcomp(datos_imputados, scale. = TRUE)


summary(pca_poliformat)
```

```
## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     6.0208 2.25020 2.09184 1.80407 1.70725 1.50712 1.2487
## Proportion of Variance 0.5254 0.07338 0.06342 0.04717 0.04224 0.03292 0.0226
## Cumulative Proportion  0.5254 0.59874 0.66216 0.70933 0.75157 0.78449 0.8071
##                           PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     1.14679 1.04552 0.98868 0.92634 0.88925 0.86480 0.83107
## Proportion of Variance 0.01906 0.01584 0.01417 0.01244 0.01146 0.01084 0.01001
## Cumulative Proportion  0.82615 0.84199 0.85616 0.86860 0.88006 0.89089 0.90090
##                           PC15    PC16    PC17    PC18   PC19   PC20    PC21
## Standard deviation     0.79793 0.77408 0.75903 0.70587 0.6848 0.6699 0.62896
## Proportion of Variance 0.00923 0.00868 0.00835 0.00722 0.0068 0.0065 0.00573
## Cumulative Proportion  0.91013 0.91882 0.92717 0.93439 0.9412 0.9477 0.95342
##                           PC22    PC23    PC24    PC25    PC26    PC27    PC28
## Standard deviation     0.56012 0.53407 0.49265 0.48928 0.45595 0.43767 0.40262
```

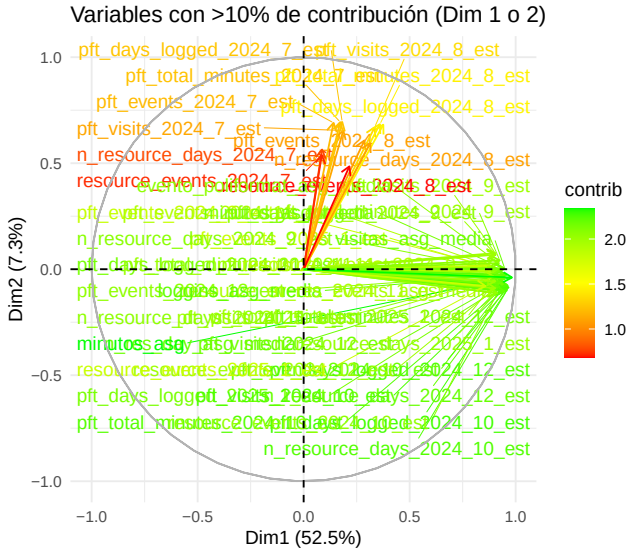
```
## Proportion of Variance 0.00455 0.00413 0.00352 0.00347 0.00301 0.00278 0.00235
## Cumulative Proportion  0.95797 0.96210 0.96562 0.96909 0.97210 0.97488 0.97722
##                           PC29    PC30    PC31    PC32    PC33    PC34    PC35
## Standard deviation     0.37913 0.36678 0.33693 0.32676 0.31526 0.30520 0.29679
## Proportion of Variance 0.00208 0.00195 0.00165 0.00155 0.00144 0.00135 0.00128
## Cumulative Proportion  0.97931 0.98126 0.98290 0.98445 0.98589 0.98724 0.98852
##                           PC36    PC37    PC38    PC39    PC40    PC41    PC42
## Standard deviation     0.25831 0.25063 0.24768 0.23638 0.22778 0.22323 0.20434
## Proportion of Variance 0.00097 0.00091 0.00089 0.00081 0.00075 0.00072 0.00061
## Cumulative Proportion  0.98948 0.99039 0.99128 0.99209 0.99285 0.99357 0.99417
##                           PC43    PC44    PC45    PC46    PC47    PC48    PC49
## Standard deviation     0.19514 0.19236 0.18753 0.18037 0.17343 0.16764 0.16221
## Proportion of Variance 0.00055 0.00054 0.00051 0.00047 0.00044 0.00041 0.00038
## Cumulative Proportion  0.99472 0.99526 0.99577 0.99624 0.99668 0.99709 0.99747
##                           PC50    PC51    PC52    PC53    PC54    PC55    PC56
## Standard deviation     0.15088 0.13739 0.12945 0.12543 0.11475 0.11227 0.10620
## Proportion of Variance 0.00033 0.00027 0.00024 0.00023 0.00019 0.00018 0.00016
## Cumulative Proportion  0.99780 0.99807 0.99831 0.99854 0.99873 0.99891 0.99908
##                           PC57    PC58    PC59    PC60    PC61    PC62    PC63
## Standard deviation     0.10170 0.09382 0.08866 0.08643 0.07645 0.07204 0.07146
## Proportion of Variance 0.00015 0.00013 0.00011 0.00011 0.00008 0.00008 0.00007
## Cumulative Proportion  0.99923 0.99936 0.99947 0.99958 0.99966 0.99974 0.99981
##                           PC64    PC65    PC66    PC67    PC68      PC69
## Standard deviation     0.06965 0.06433 0.05451 0.02535 0.01934 7.547e-15
## Proportion of Variance 0.00007 0.00006 0.00004 0.00001 0.00001 0.000e+00
## Cumulative Proportion  0.99988 0.99994 0.99999 0.99999 1.00000 1.000e+00
```

```r
fviz_eig(pca_poliformat, addlabels = TRUE, main="PCA Poliformat")
```

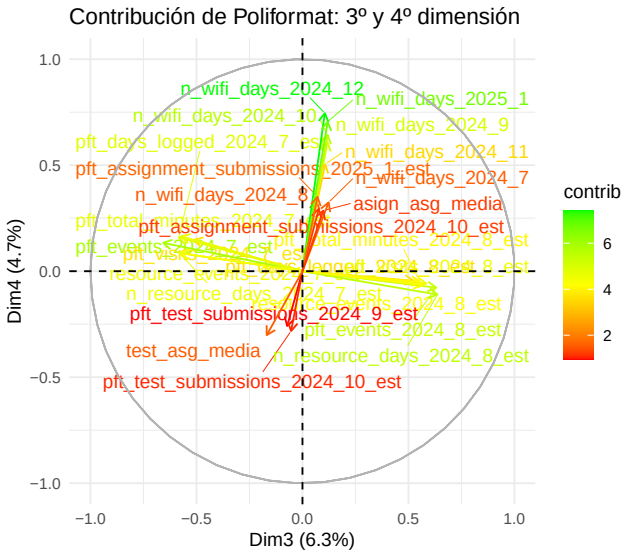



```r
contrib=factoextra::get_pca_var(pca_poliformat)$contrib



var_filtradas=rownames(contrib)[apply(contrib[, 1:2], 1, function(x) any(x > 2))]
```

```
var_filtradas2=rownames(contrib)[apply(contrib[, 3:4], 1, function(x) any(x > 2))]


fviz_pca_var(pca_poliformat,
             select.var = list(name = var_filtradas),
             col.var = "contrib",
             gradient.cols = c("red", "yellow", "green"),
             repel = TRUE) +
  labs(title = "Variables con >10% de contribución (Dim 1 o 2)") +
  theme_minimal()
```




```
fviz_pca_var(pca_poliformat,
             axes=c(3,4),
             select.var = list(name=var_filtradas2),
             col.var = "contrib",
             gradient.cols = c("red", "yellow", "green"),
             repel = TRUE) +
  labs(title = "Contribución de Poliformat: 3º y 4º dimensión") +
  theme_minimal()
```



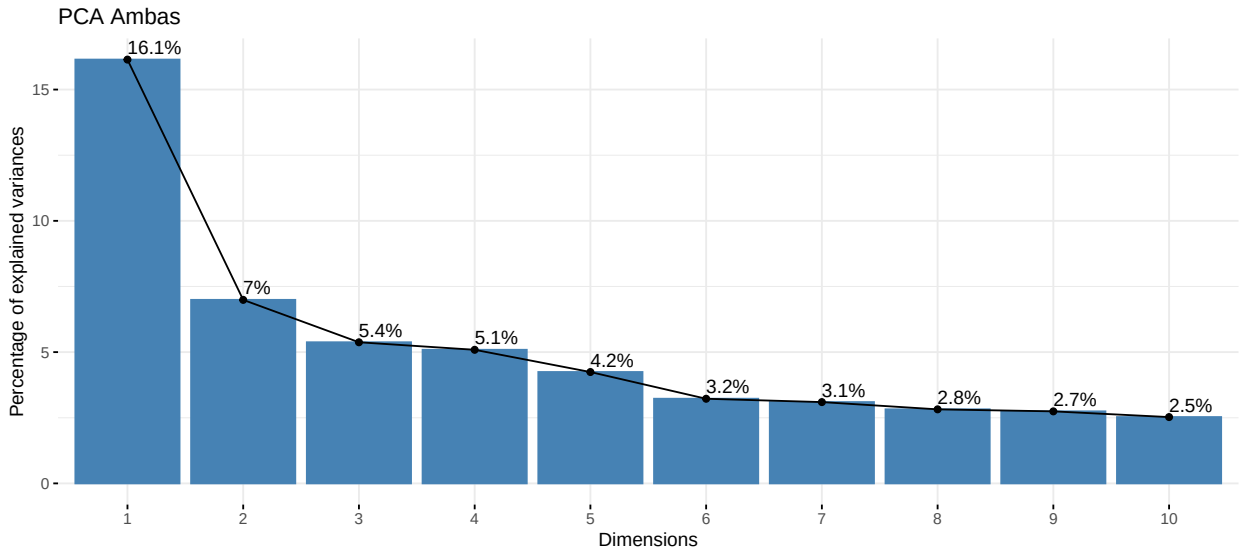
Academicas + sociodemografía

```
ambas= academicas %>%
  left_join(select(sociodemografia,-abandono),by="dni_hash")

ambas
```

```
## # A tibble: 1,754 x 59
##    dni_hash     curso_mas_bajo curso_mas_alto cred_mat1 cred_mat2 cred_mat3
##    <chr>        <fct>          <fct>               <dbl>     <dbl>     <dbl>
##  1 000d5aac651d 2              3                       0       9        15
##  2 001735e5198e 2              2                       0      60         0
##  3 004f95523855 3              4                       0       0        36
##  4 005ae39cc97a 2              4                       0      12        22.5
##  5 0062a463b0c1 2              3                       0      15        45
##  6 008e8f24466f 3              3                       0       0        60
##  7 00d278569c9f 1              1                      60       0         0
##  8 00dc0e8a83fd 4              4                       0       0         0
##  9 0113867b543d 1              1                      12      49.5       0
## 10 014417911743 2              2                       0      60         0
## # i 1,744 more rows
## # i 53 more variables: cred_mat4 <dbl>, cred_sup_normal <dbl>,
## #   cred_sup_espec <dbl>, cred_sup <dbl>, cred_mat_normal <dbl>,
## #   cred_mat_movilidad <dbl>, cred_ptes_acta <dbl>, cred_mat_practicas <dbl>,
## #   cred_mat_sem_a <dbl>, cred_mat_sem_b <dbl>, cred_mat_anu <dbl>,
## #   cred_mat_total <dbl>, cred_sup_sem_a <dbl>, cred_sup_sem_b <dbl>,
## #   cred_sup_total <dbl>, rendimiento_cuat_a <dbl>, ...
```

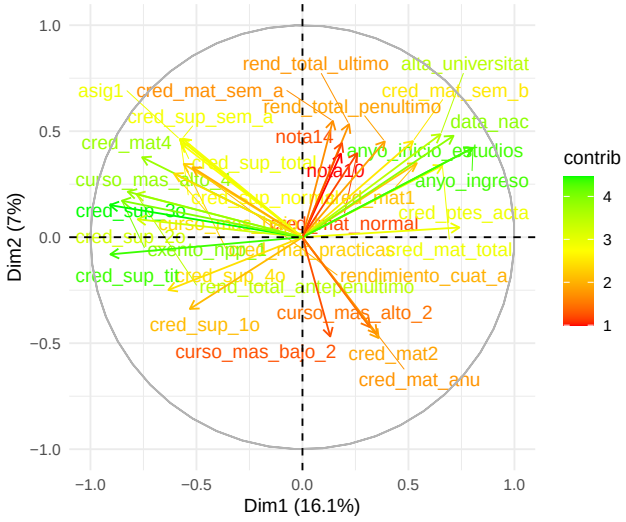
```
num_vars=ambas %>% select(where(is.numeric))
cat_vars=ambas %>% select(where(is.factor) | where(is.character)) %>%
  select(-dni_hash)


cat_dummies=dummy_cols(cat_vars, remove_first_dummy = TRUE, remove_selected_columns = TRUE)
```

```r
datos_pca=bind_cols(num_vars, cat_dummies)


datos_pca=datos_pca[, sapply(datos_pca, function(x) var(x, na.rm = TRUE) > 0)]


preprocess=preProcess(datos_pca, method = "medianImpute")
datos_imputados=predict(preprocess, newdata = datos_pca)


pca_resultado=prcomp(datos_imputados, scale. = TRUE)

contrib=factoextra::get_pca_var(pca_resultado)$contrib
var_filtradas=rownames(contrib)[apply(contrib[, 1:2], 1, function(x) any(x > 2))]
var_filtradas2=rownames(contrib)[apply(contrib[, 3:4], 1, function(x) any(x > 2))]

fviz_eig(pca_resultado, addlabels = TRUE,main="PCA Ambas")
```



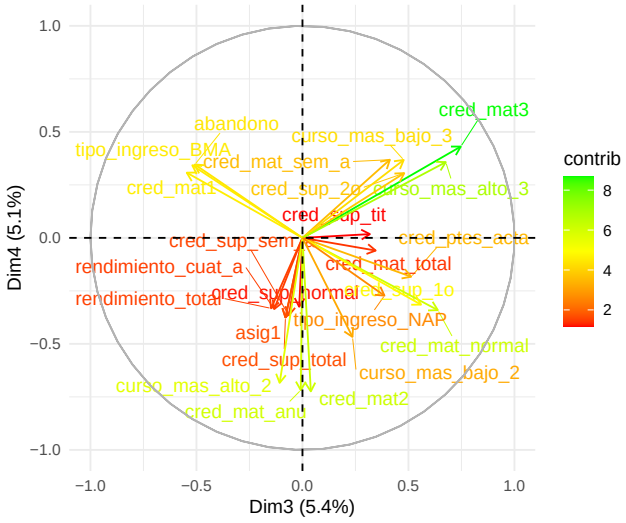

```r
fviz_pca_var(pca_resultado,
             select.var = list(name = var_filtradas),
             col.var = "contrib",
             gradient.cols = c("red", "yellow", "green"),
             repel = TRUE) +
  labs(title = "Contribución de variables académicas y sociodemográficas: 1º y 2º dimensión") +
  theme_minimal()
```

Contribución de variables académicas y sociodemográficas: 1º y 2º dimensión



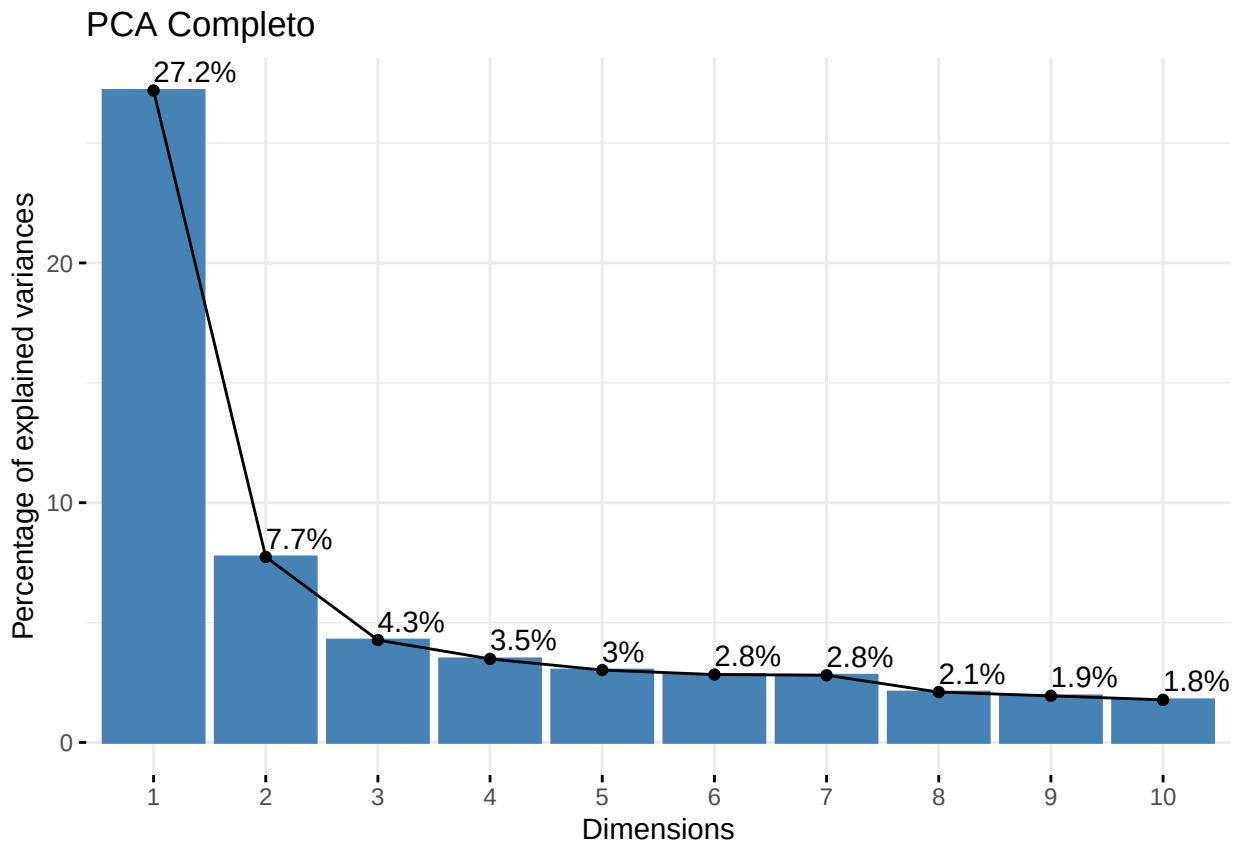

```
fviz_pca_var(pca_resultado,
             axes=c(3,4),
             select.var = list(name=var_filtradas2),
             col.var = "contrib",
             gradient.cols = c("red", "yellow", "green"),
             repel = TRUE) +
  labs(title = "Contribución de variables académicas y sociodemográficas: 3º y 4º dimensión") +
  theme_minimal()
```

Contribución de variables académicas y sociodemográficas: 3º y 4º dimensión

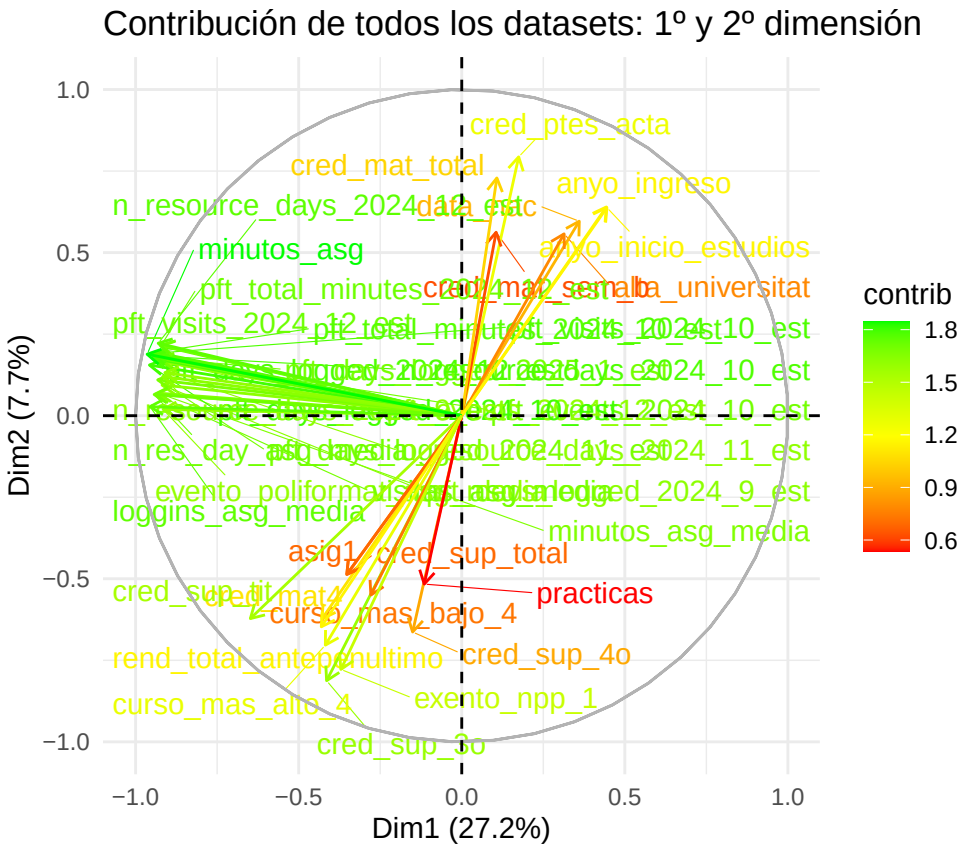



```
completo=sociodemografia %>%
  left_join(select(academicas,-c("abandono")), by = "dni_hash") %>%
  left_join(select(poliformat,-c("abandono")), by="dni_hash")



num_vars=completo %>% select(where(is.numeric))
```

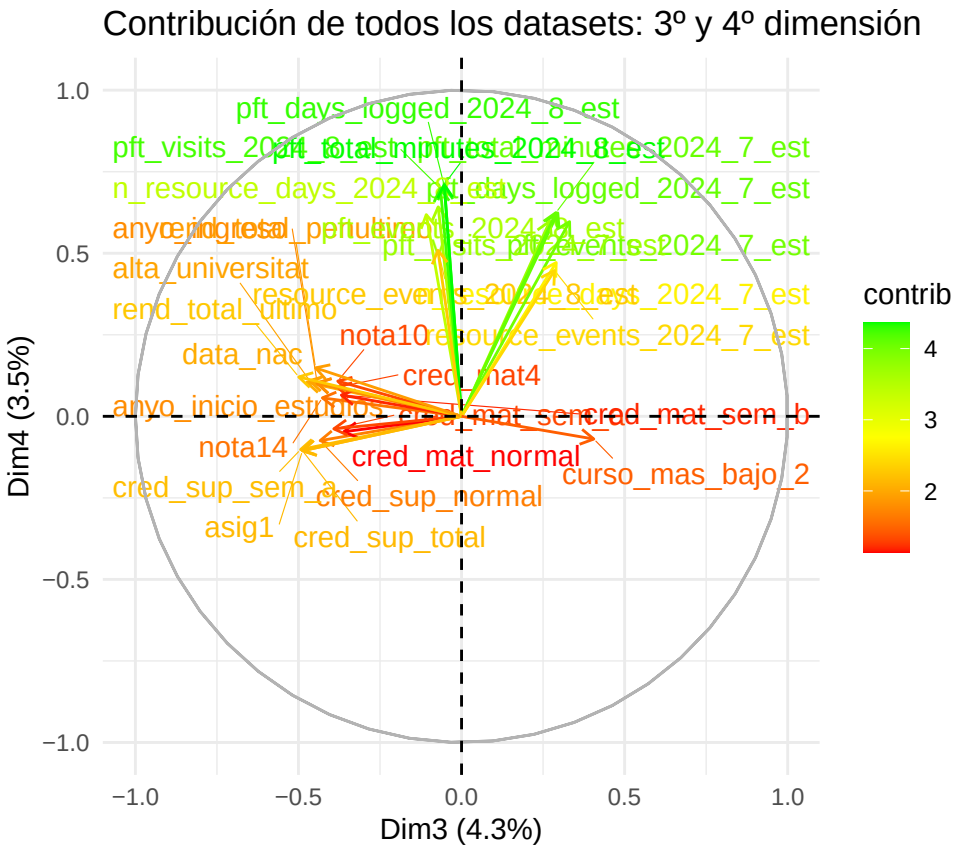
```r
cat_vars=completo %>% select(where(is.factor) | where(is.character)) %>%
  select(-dni_hash)


cat_dummies=dummy_cols(cat_vars, remove_first_dummy = TRUE, remove_selected_columns = TRUE)


datos_pca=bind_cols(num_vars, cat_dummies)


datos_pca=datos_pca[, sapply(datos_pca, function(x) var(x, na.rm = TRUE) > 0)]


preprocess=preProcess(datos_pca, method = "medianImpute")
datos_imputados=predict(preprocess, newdata = datos_pca)


pca_resultado=prcomp(datos_imputados, scale. = TRUE)
contrib=factoextra::get_pca_var(pca_resultado)$contrib
var_filtradas=rownames(contrib)[apply(contrib[, 1:2], 1, function(x) any(x > 2))]
var_filtradas2=rownames(contrib)[apply(contrib[, 3:4], 1, function(x) any(x > 2))]

fviz_eig(pca_resultado, addlabels = TRUE, main="PCA Completo")
```



```
fviz_pca_var(pca_resultado,
             select.var = list(name = var_filtradas),
             col.var = "contrib",
             gradient.cols = c("red", "yellow", "green"),
             repel = TRUE) +
  labs(title = "Contribución de todos los datasets: 1º y 2º dimensión") +
  theme_minimal()
```




```
fviz_pca_var(pca_resultado,
             axes=c(3,4),
             select.var = list(name=var_filtradas2),
             col.var = "contrib",
             gradient.cols = c("red", "yellow", "green"),
             repel = TRUE) +
  labs(title = "Contribución de todos los datasets: 3º y 4º dimensión") +
  theme_minimal()
```




```
save( academicas, sociodemografia, poliformat,abandono,file="../../Datos/Capítulos/Exploratorio.RData")
save(pca_poliformat,file="../../Datos/Capítulos/pca_poliformat.Rdata")
```