3. Caracterización del alumnado que abandona

Pablo Parrilla Cañadas

2025-04-29

Librerías

```r
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
```

```r
library(ggplot2)
```

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

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

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

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

```r
library(skimr)
```

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

```r
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(rcompanion)
library(DescTools)
```

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

```
##
## Adjuntando el paquete: 'DescTools'
```

```
## The following objects are masked from 'package:caret':
##
##     MAE, RMSE
```

```r
library(vegan)
```

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

```
## Cargando paquete requerido: permute
```

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

```
##
## Adjuntando el paquete: 'vegan'
```

```
## The following object is masked from 'package:caret':
##
##     tolerance
```

```r
library(fastDummies)
```

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

```r
library(stringr)
library(forcats)
library(DescTools)
library(patchwork)

load("../../Datos/Capítulos/Exploratorio.RData")
```

Contextualización

```r
table(abandono$mes)
```

```
##
##      enero      julio septiembre    octubre  diciembre
##          2         23         20          5          2
```

```r
abandono_date=abandono %>%
  mutate(baja_fecha = as.Date(baja_fecha))

abandono_acumulado=abandono_date[abandono_date$mes=="septiembre",] %>%
  count(baja_fecha) %>%
  arrange(baja_fecha) %>%
  mutate(acumulado = cumsum(n))


a=ggplot(abandono_acumulado, aes(x = baja_fecha, y = acumulado)) +
  geom_line(color = "black", size = 1) +
  geom_point(color = "red", size = 3) +
   scale_x_date(date_breaks = "2 days", date_labels = "%d-%m") +
   geom_text(aes(label = acumulado), vjust = -1, hjust= 1, color = "black", size = 3) +
  labs(title = "Bajas Acumuladas por día de septiembre",
       x = "Fecha",
       y = "Total Acumulado") +
  theme_minimal()+
   theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## 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
abandono_acumulado=abandono_date[abandono_date$mes=="julio",] %>%
  count(baja_fecha) %>%
  arrange(baja_fecha) %>%
  mutate(acumulado = cumsum(n))


b=ggplot(abandono_acumulado, aes(x = baja_fecha, y = acumulado)) +
  geom_line(color = "black", size = 1) +
  geom_point(color = "red", size = 3) +
   scale_x_date(date_breaks = "2 days", date_labels = "%d-%m", position = ) +
   geom_text(aes(label = acumulado), vjust = -1, hjust=0.75, color = "black", size = 3) +
  labs(title = "Bajas Acumuladas por día de julio",
       x = "Fecha",
       y = "Total Acumulado") +
  theme_minimal()+
   theme(axis.text.x = element_text(angle = 45, hjust = 1))

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

Bajas Acumuladas por día de julio

Bajas Acumuladas por día de septiembre
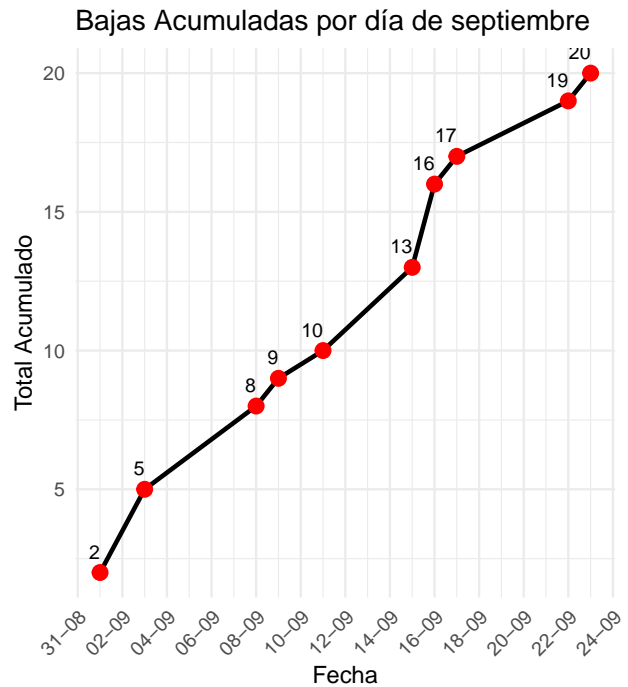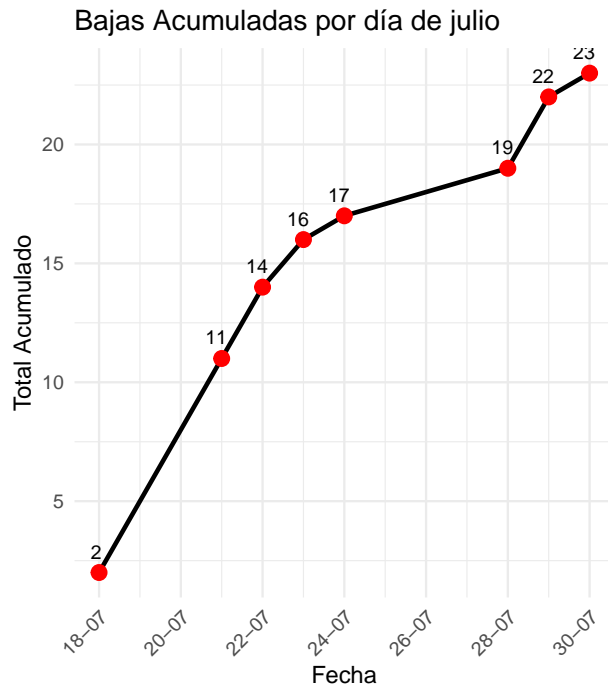
```r
total_abandonadas=sum(abandono$asi_left, na.rm = TRUE)


media_abandono_por_mes=abandono %>%
  group_by(mes) %>%
  summarise(
    media_asi_left = mean(asi_left, na.rm = TRUE),
    porcentaje = (sum(asi_left, na.rm = TRUE) / total_abandonadas) * 100
  ) %>% arrange(desc(porcentaje))



print(media_abandono_por_mes)
```

```
## # A tibble: 5 x 3
##   mes         media_asi_left porcentaje
##   <fct>                <dbl>      <dbl>
## 1 julio                   10       44.1
## 2 septiembre            10.4       40.0
## 3 octubre                  9        8.62
## 4 diciembre               11        4.21
## 5 enero                    8        3.07
```

Influencia sociodemográfica

```r
vars_usar=c("nacionalitat", "data_nac", "sexe", "alta_universitat", "prov_origen",
            "anyo_ingreso", "tipo_ingreso", "nota10", "nota14", "estudios_p", "estudios_m",
            "dedicacion", "desplazado", "discapacidad", "becado", "preferencia_seleccion", "abandono
```

```r
df_sel=sociodemografia[, vars_usar]

df_sel$abandono=as.factor(df_sel$abandono)




df_sel$abandono=as.numeric(as.character(df_sel$abandono))


num_vars=vars_usar[sapply(df_sel[, vars_usar], is.numeric)]
cat_vars=vars_usar[sapply(df_sel[, vars_usar], function(x) is.character(x) || is.factor(x))]


resultados=data.frame(variable = character(),
                      correlacion = numeric(),
                      tipo = character(),
                      stringsAsFactors = FALSE)


for (var in num_vars) {
  if (var != "abandono") {
    valor=suppressWarnings(cor(df_sel$abandono, df_sel[[var]], use = "complete.obs"))
    resultados=rbind(resultados, data.frame(variable = var, correlacion = valor, tipo = "Numérica"))
  }
}


for (var in cat_vars) {
  x=as.factor(df_sel$abandono)
  y=as.factor(df_sel[[var]])

  df_temp=data.frame(x = x, y = y)
  df_temp=df_temp[complete.cases(df_temp), ]


  if (nlevels(df_temp$x) >= 2 && nlevels(df_temp$y) >= 2) {

    df_temp$y=fct_lump_min(df_temp$y, min = 2)
    df_temp$y=droplevels(df_temp$y)


    v=DescTools::CramerV(table(df_temp$x, df_temp$y), bias.correct = TRUE)

    resultados=rbind(resultados, data.frame(variable = var, correlacion = v, tipo = "Categórica"))
  }
}


head(resultados)
```

```
##           variable  correlacion       tipo
```

```
## 1        data_nac  0.045453781    Numérica
## 2 alta_universitat  0.049245330    Numérica
## 3     anyo_ingreso  0.094659131    Numérica
## 4           nota10 -0.013135812    Numérica
## 5           nota14 -0.009261598    Numérica
## 6      nacionalitat  0.034053454 Categórica
```

```r
ggplot(resultados, aes(x = reorder(variable, correlacion), y = correlacion, fill = tipo)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  scale_fill_manual(values = c("Numérica" = "steelblue", "Categórica" = "orange")) +
  theme_minimal() +
  labs(title = "Correlación de variables sociodemográficas con 'abandono'",
       x = "Variable",
       y = "Correlación (Pearson o Cramér's V)",
       fill = "Tipo de variable")
```


Correlación de variables sociodemográficas con 'abandono'
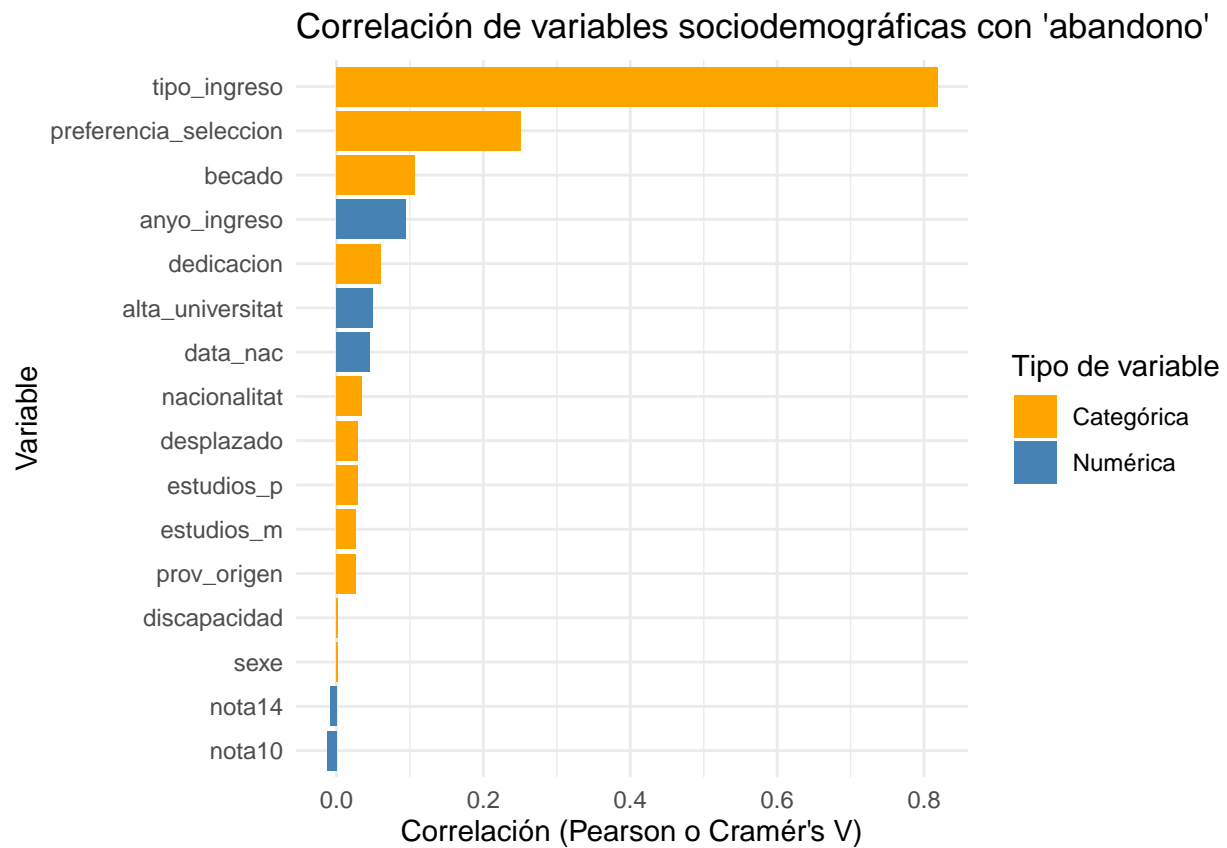
```r
table(sociodemografia$preferencia_seleccion, sociodemografia$abandono)
```

```
##
##                0    1 Desconocido
##   1         1250   17           0
##   2          230    5           0
##   3           99   12           0
##   Baja        69   16           0
##   Desconocido 54    2           0
```

```r
table(sociodemografia$becado,sociodemografia$abandono)
```

```
##
##                 0    1 Desconocido
##   0          1181   51           0
##   1            33    0           0
##   2           488    1           0
##   Desconocido   0    0           0
```

```r
df_long=sociodemografia[sociodemografia$abandono==1,] %>%
  pivot_longer(cols = c(anyo_ingreso, alta_universitat,data_nac), names_to = "variable", values_to = "va

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()
```
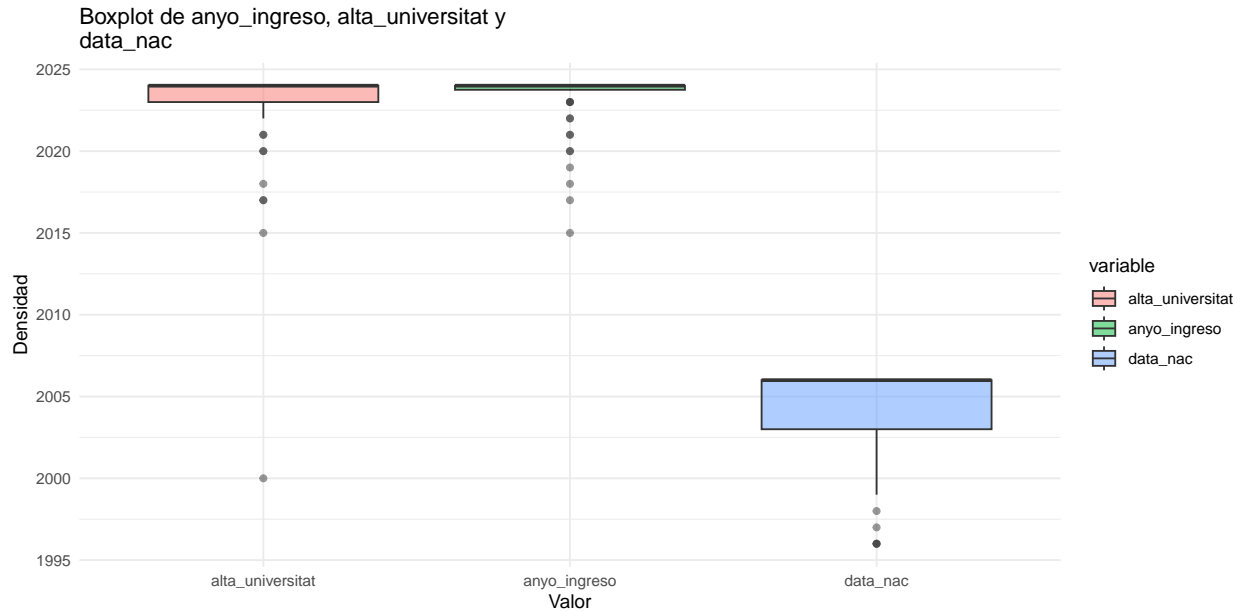


```r
multi_vars=c("tipo_ingreso", "preferencia_seleccion","becado", "estudios_p", "estudios_m")

plots_multi=lapply(multi_vars, function(var) {
  ggplot(sociodemografia[sociodemografia$abandono==1,], aes(x = fct_infreq(.data[[var]]), fill = .data[
    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)
```

tipo_ingreso

preferencia_seleccion

becado

estudios_p

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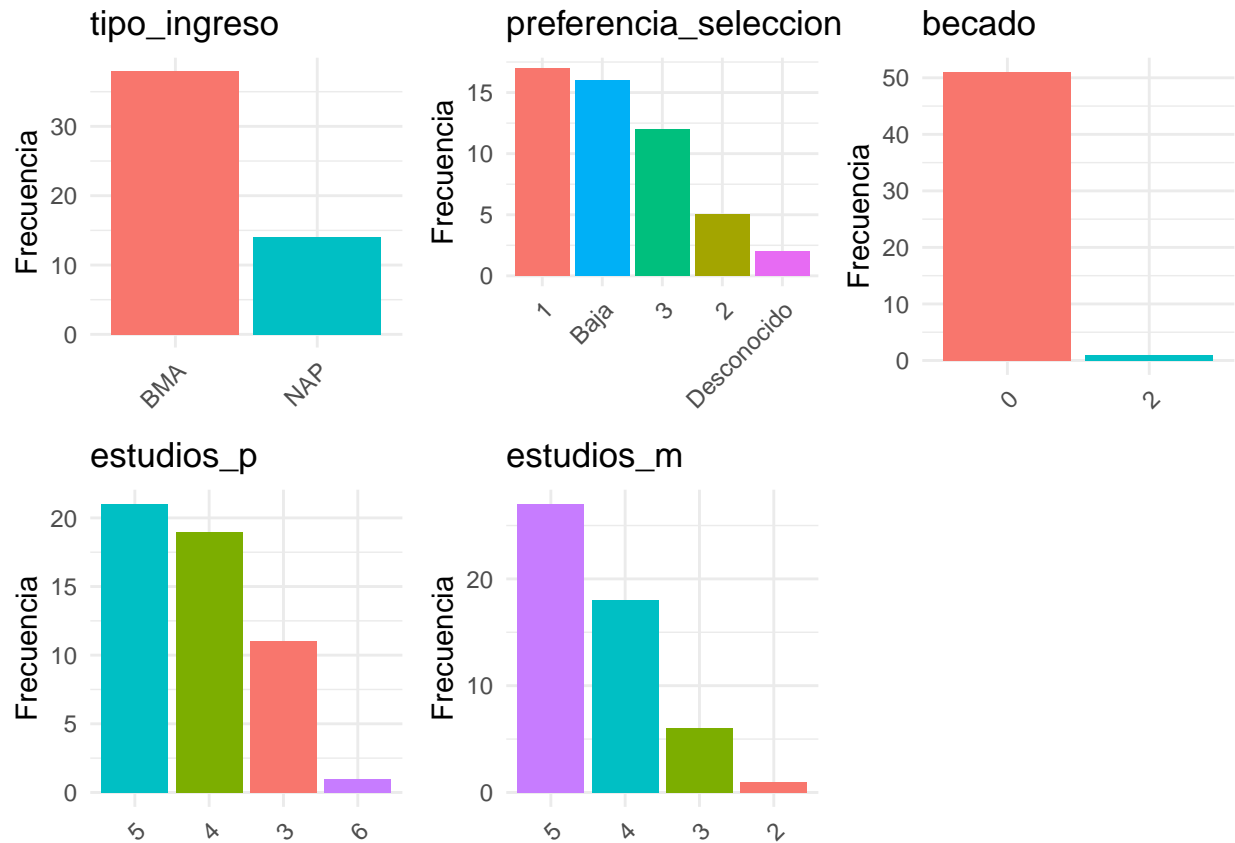
```r
binarias_1=c("nacionalitat", "sexe")
binarias_3=c("prov_origen", "desplazado")
binarias_2=c("discapacidad","dedicacion" )


crear_plot_binarias=function(vars) {
  df=sociodemografia[sociodemografia$abandono==1,] %>%
    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 = 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",
```
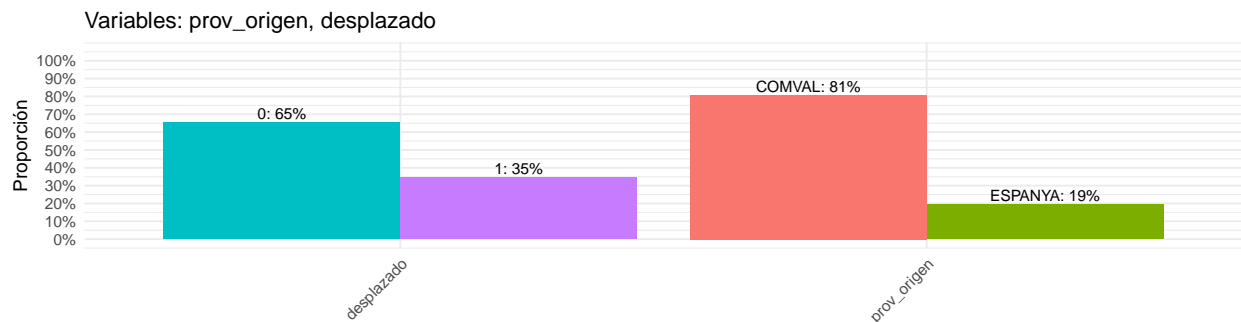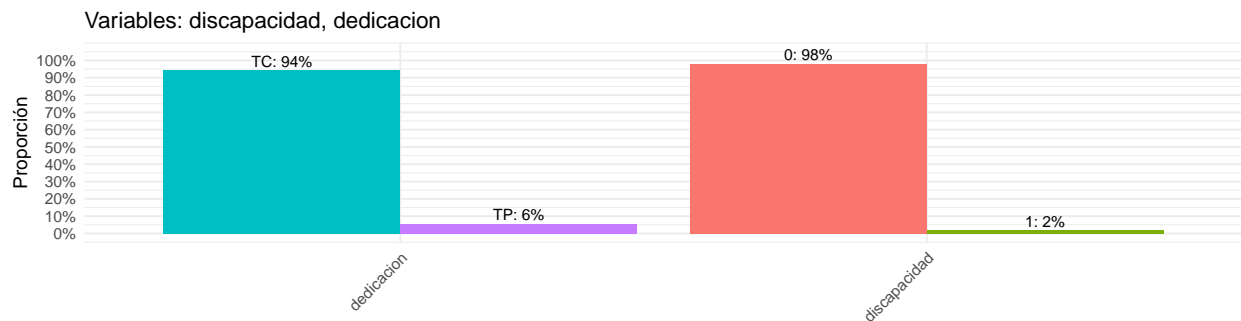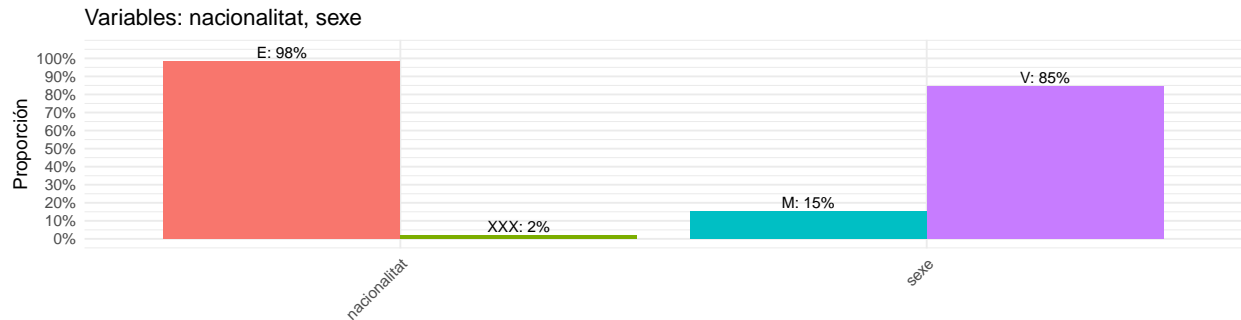
```
            axis.text.x = element_text(angle = 45, hjust = 1))
}


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)
```

Variables: nacionalitat, sexe

Variables: discapacidad, dedicacion

Variables: prov_origen, desplazado

```
vars=c("nacionalitat", "data_nac", "alta_universitat", "prov_origen",
               "anyo_ingreso", "tipo_ingreso",
               "dedicacion", "desplazado", "becado", "preferencia_seleccion", "abandono")


df_model=sociodemografia[, c("abandono", vars)] %>% na.omit()


df_encoded=dummy_cols(df_model, select_columns = c("tipo_ingreso","dedicacion", "desplazado", "becado",
                      remove_selected_columns = TRUE, remove_first_dummy = FALSE)
```

```r
dist_mat=dist(df_encoded %>% select(-abandono), method = "euclidean")
```

```
## Warning in dist(df_encoded %>% select(-abandono), method = "euclidean"): NAs
## introducidos por coerción
```

```r
a=adonis2(dist_mat ~ abandono, data = df_encoded, permutations = 999)
a
```

```
## Permutation test for adonis under reduced model
## Permutation: free
## Number of permutations: 999
##
## adonis2(formula = dist_mat ~ abandono, data = df_encoded, permutations = 999)
##            Df SumOfSqs      R2      F Pr(>F)
## Model       1      253 0.00719 12.643  0.002 **
## Residual 1747    35011 0.99281
## Total    1748    35265 1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
vars_usar=colnames(academicas[2:43])



df_sel=academicas[, vars_usar]

df_sel$abandono=as.factor(df_sel$abandono)




df_sel$abandono=as.numeric(as.character(df_sel$abandono))


num_vars=vars_usar[sapply(df_sel[, vars_usar], is.numeric)]
cat_vars=vars_usar[sapply(df_sel[, vars_usar], function(x) is.character(x) || is.factor(x))]


resultados=data.frame(variable = character(),
                      correlacion = numeric(),
                      tipo = character(),
                      stringsAsFactors = FALSE)


for (var in num_vars) {
  if (var != "abandono") {
    valor=suppressWarnings(cor(df_sel$abandono, df_sel[[var]], use = "complete.obs"))
    resultados=rbind(resultados, data.frame(variable = var, correlacion = valor, tipo = "Numérica"))
  }
}
```

```r
for (var in cat_vars) {
  x=as.factor(df_sel$abandono)
  y=as.factor(df_sel[[var]])

  df_temp=data.frame(x = x, y = y)
  df_temp=df_temp[complete.cases(df_temp), ]


  if (nlevels(df_temp$x) >= 2 && nlevels(df_temp$y) >= 2) {

    df_temp$y=fct_lump_min(df_temp$y, min = 2)
    df_temp$y=droplevels(df_temp$y)


    v=DescTools::CramerV(table(df_temp$x, df_temp$y), bias.correct = TRUE)

    resultados=rbind(resultados, data.frame(variable = var, correlacion = v, tipo = "Categórica"))
  }
}


resultados
```

```
##                      variable  correlacion        tipo
## 1                    cred_mat1  0.200996391    Numérica
## 2                    cred_mat2 -0.068895342    Numérica
## 3                    cred_mat3 -0.054792806    Numérica
## 4                    cred_mat4 -0.085974714    Numérica
## 5               cred_sup_normal -0.067834838    Numérica
## 6               cred_sup_espec  0.019398897    Numérica
## 7                     cred_sup -0.003629293    Numérica
## 8             cred_mat_normal -0.645019286    Numérica
## 9           cred_mat_movilidad -0.025934414    Numérica
## 10             cred_ptes_acta -0.466742298    Numérica
## 11          cred_mat_practicas -0.075412629    Numérica
## 12             cred_mat_sem_a  0.045173808    Numérica
## 13             cred_mat_sem_b  0.098442042    Numérica
## 14               cred_mat_anu -0.052688160    Numérica
## 15              cred_mat_total  0.057849755    Numérica
## 16              cred_sup_sem_a -0.052354461    Numérica
## 17              cred_sup_sem_b -0.005905679    Numérica
## 18              cred_sup_total -0.052615682    Numérica
## 19          rendimiento_cuat_a -0.010423215    Numérica
## 20           rendimiento_total -0.009281606    Numérica
## 21      anyo_inicio_estudios  0.094010916    Numérica
## 22                 cred_sup_1o -0.197943323    Numérica
## 23                 cred_sup_2o -0.121173923    Numérica
## 24                 cred_sup_3o -0.084125482    Numérica
## 25                 cred_sup_4o -0.033958947    Numérica
## 26                   practicas -0.021604612    Numérica
## 27                 actividades -0.024961149    Numérica
## 28                      ajuste -0.064502312    Numérica
## 29               cred_sup_tit -0.151681689    Numérica
```

```
## 30                    asig1 -0.052615682   Numérica
## 31                    pract1          NA   Numérica
## 32                    activ1          NA   Numérica
## 33                    total1          NA   Numérica
## 34         rend_total_ultimo -0.159965203   Numérica
## 35      rend_total_penultimo -0.094552718   Numérica
## 36 rend_total_antepenultimo -0.061751416   Numérica
## 37           curso_mas_bajo  0.229612344 Categórica
## 38           curso_mas_alto  0.273260245 Categórica
## 39               exento_npp  0.097463163 Categórica
## 40             es_retitulado  0.005905679 Categórica
## 41               es_adaptado  0.047330105 Categórica
```

```
ggplot(resultados, aes(x = reorder(variable, correlacion), y = correlacion, fill = tipo)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  scale_fill_manual(values = c("Numérica" = "steelblue", "Categórica" = "orange")) +
  theme_minimal() +
  theme(axis.text.y=element_text(size = 6))+
  labs(title = "Correlación de variables académicas con 'abandono'",
       x = "Variable",
       y = "Correlación (Pearson o Cramér's V)",
       fill = "Tipo de variable")
```

```
## Warning: Removed 3 rows containing missing values or values outside the scale range
## ('geom_bar()').
```


Correlación de variables académicas con 'abandono'
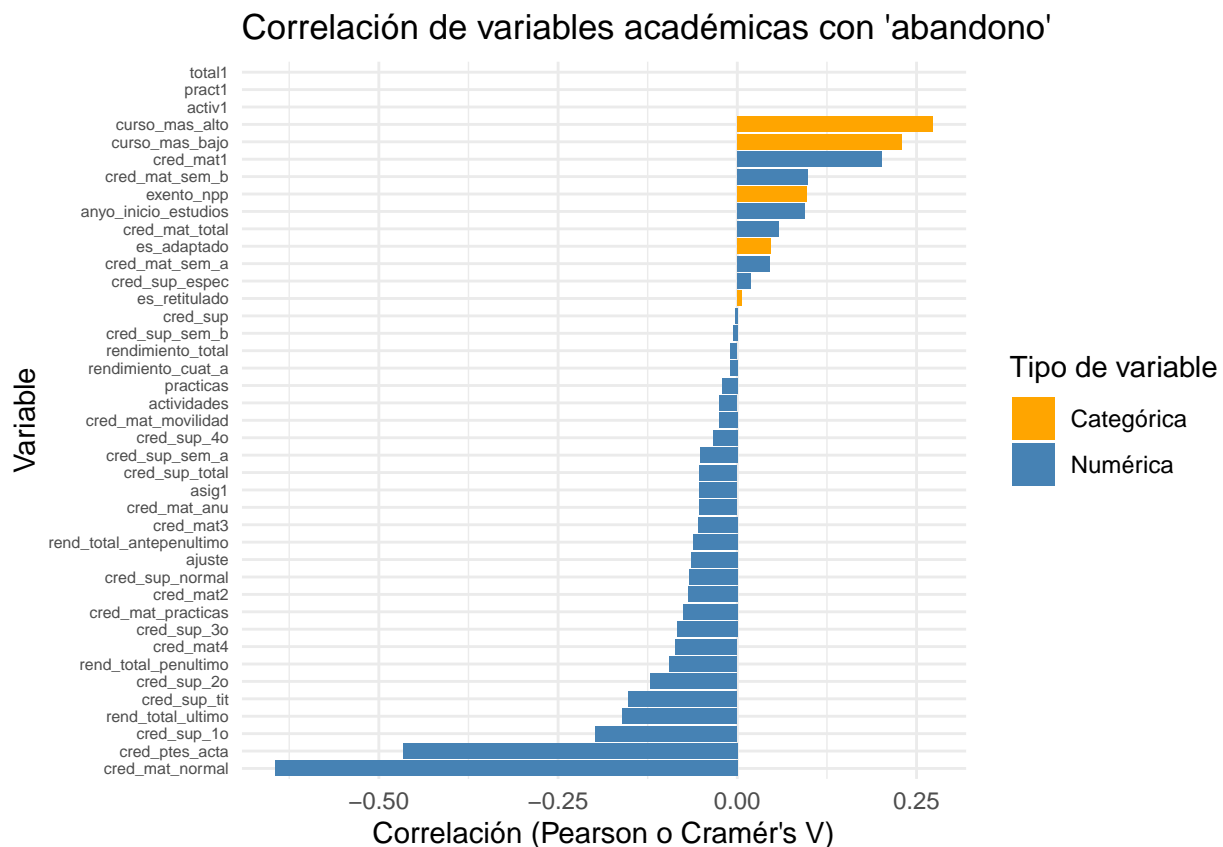
```r
table(academicas$curso_mas_bajo,academicas$curso_mas_alto,academicas$abandono)
```

```
## , ,  = 0
##
##
##       1   2   3   4
##   1 388 138  16   4
##   2   0 270 191  89
##   3   0   0 189 160
##   4   0   0   0 257
##
## , ,  = 1
##
##
##       1   2   3   4
##   1  48   0   2   0
##   2   0   1   1   0
##   3   0   0   0   0
##   4   0   0   0   0
```

```r
grupo_creditos_altos=c( "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",   "c:

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[academicas$abandono==1,] %>%
  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 de personas ", x = "Valor", y = "Variable") +
  theme_minimal(base_size = 12)
```
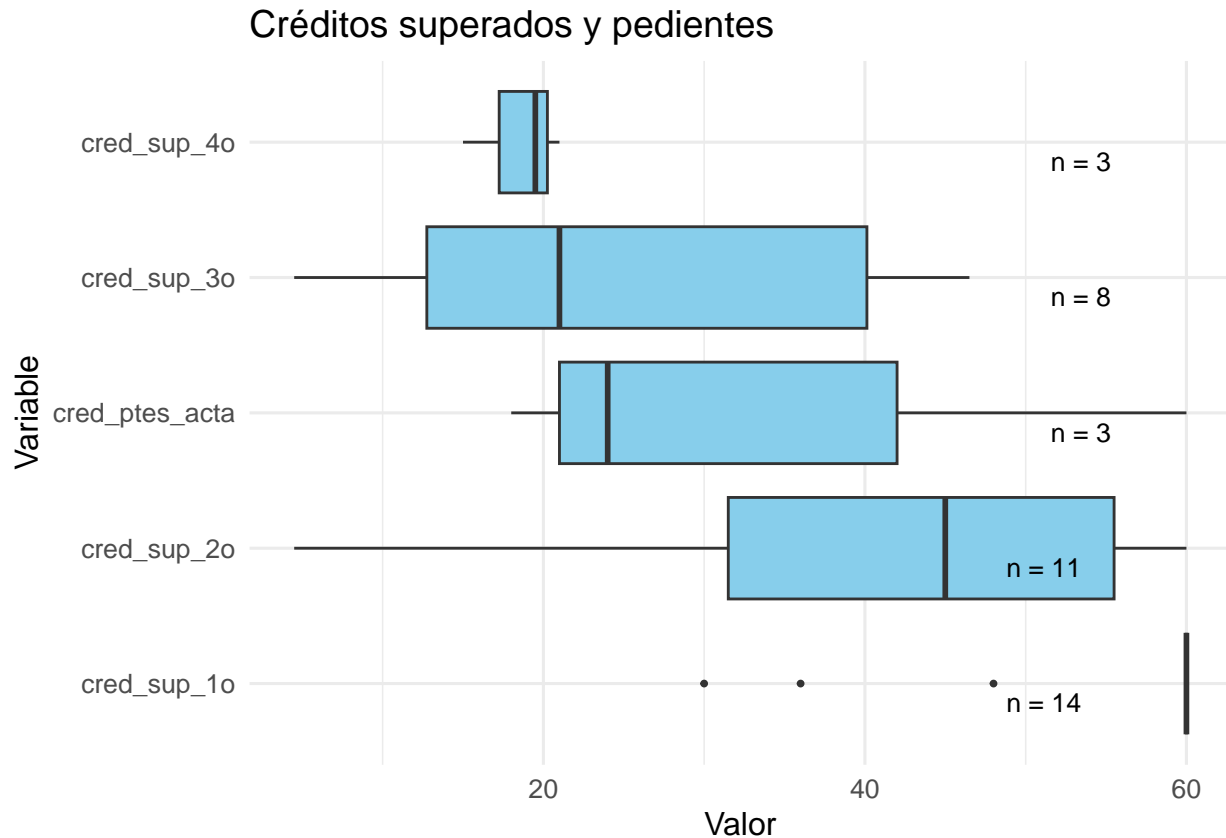
Distribución de créditos totales de personas
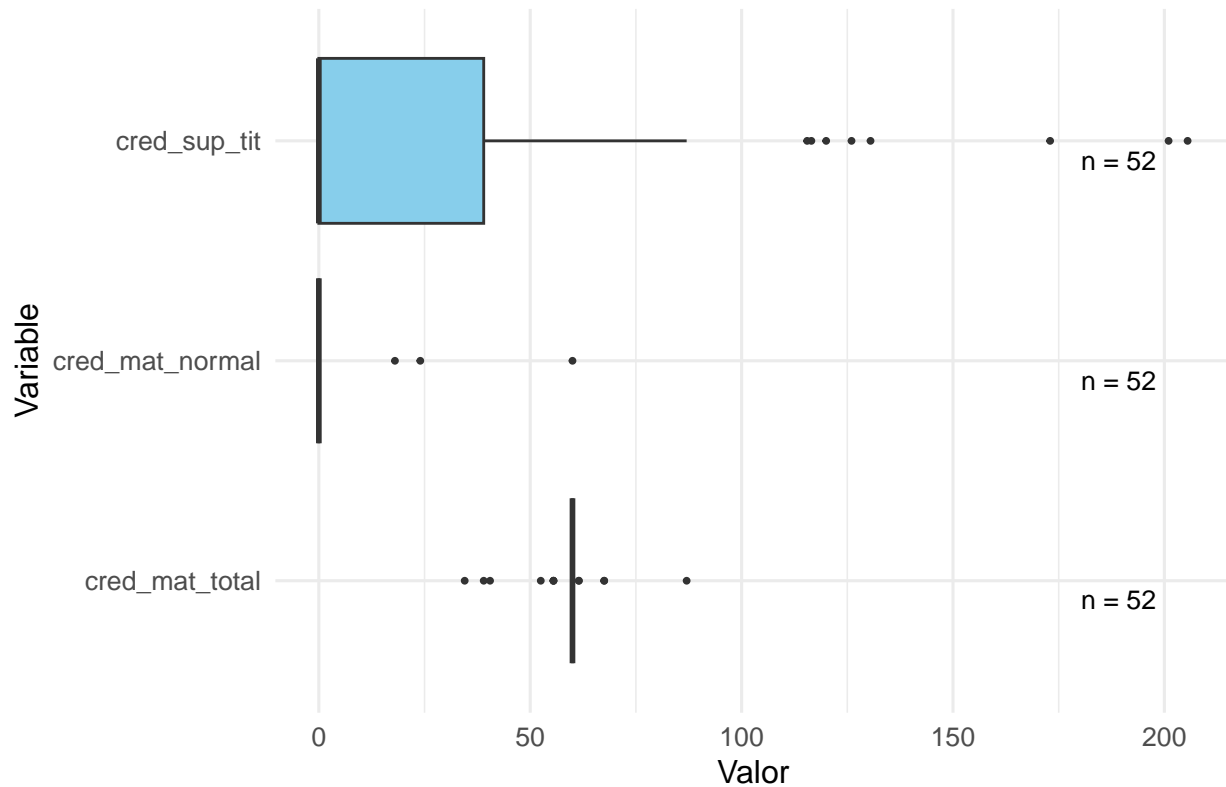

```r
df_long=academicas[academicas$abandono==1,] %>%
  select(all_of(grupo_creditos_moderados))

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 221 missing values.
## i Use `.na_rm = TRUE` to silence this message.
## i Use `.na_rm = FALSE` to preserve NAs.

## Warning: Removed 221 rows containing non-finite outside the scale range
```

(`stat_boxplot()`).

Créditos superados y pedientes


```r
df_long=academicas[academicas$abandono==1,] %>%
  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 185 missing values.
## i Use `.na_rm = TRUE` to silence this message.
## i Use `.na_rm = FALSE` to preserve NAs.
```

```
## Warning: Removed 185 rows containing non-finite outside the scale range
## ('stat_boxplot()').
```

Créditos matriculados por año y semestre
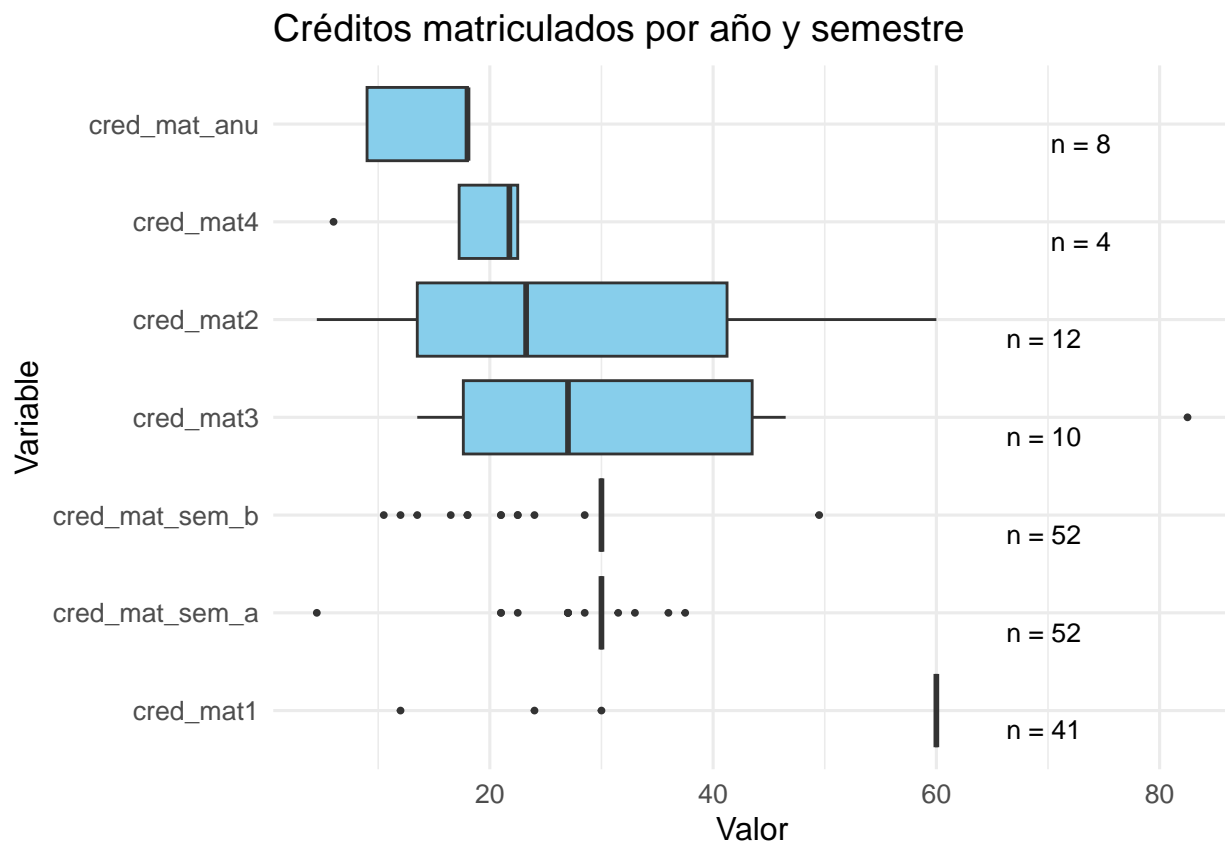

```
vars=c( "curso_mas_bajo", "curso_mas_alto", "cred_mat1",
  "cred_mat2", "cred_mat3", "cred_mat4", "cred_sup_normal",
  "cred_sup_espec", "cred_sup", "cred_mat_normal", "cred_mat_movilidad",
  "cred_ptes_acta", "cred_mat_practicas", "cred_mat_sem_a", "cred_mat_sem_b",
  "cred_mat_anu", "cred_mat_total", "cred_sup_sem_a", "cred_sup_sem_b",
  "cred_sup_total", "rendimiento_cuat_a", "rendimiento_total", "exento_npp",
  "anyo_inicio_estudios",  "cred_sup_1o",
  "cred_sup_2o", "cred_sup_3o", "cred_sup_4o", "practicas",
  "actividades", "ajuste", "cred_sup_tit", "asig1",
  "pract1", "activ1", "total1", "rend_total_ultimo",
  "rend_total_penultimo", "rend_total_antepenultimo"
)



df_model=academicas[, c("abandono", vars)]



dist_mat=dist(df_model %>% select(-abandono), method = "euclidean")
```

```
a=adonis2(dist_mat ~ abandono, data = df_model, permutations = 999)
a
```

```
## Permutation test for adonis under reduced model
## Permutation: free
## Number of permutations: 999
##
## adonis2(formula = dist_mat ~ abandono, data = df_model, permutations = 999)
##            Df SumOfSqs      R2      F Pr(>F)
## Model       1   875373 0.03245 58.755  0.001 ***
## Residual 1752 26102606 0.96755
## Total    1753 26977979 1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
vars_usar=colnames(poliformat[2:69])
```

```r
df_sel=poliformat[, vars_usar]
df_sel$abandono=as.numeric(as.character(poliformat$abandono))
```

```r
num_vars=vars_usar[sapply(df_sel[, vars_usar], is.numeric)]
cat_vars=vars_usar[sapply(df_sel[, vars_usar], function(x) is.character(x) || is.factor(x))]
```

```r
resultados=data.frame(variable = character(),
                      correlacion = numeric(),
                      tipo = character(),
                      mes = character(),
                      stringsAsFactors = FALSE)

extraer_mes=function(nombre) {

  if (str_detect(nombre, "_\\d{1,2}_est$")) {
    return(str_extract(nombre, "\\d{1,2}(?=_est$)"))
  }

  if (str_detect(nombre, "^n_wifi_days_2024_\\d{1,2}$")) {
    return(str_extract(nombre, "\\d{1,2}$"))
  }

  if (str_detect(nombre, "^n_wifi_days_2025_\\d{1,2}$")) {
    return(str_extract(nombre, "\\d{1,2}$"))
  }

  if (str_ends(nombre, "_media")) {
    return("Media")
  }
  return("Desconocido")
}

for (var in num_vars) {
```

```r
  if (var != "abandono") {
    valor=suppressWarnings(cor(df_sel$abandono, df_sel[[var]], use = "complete.obs"))
    mes=extraer_mes(var)
    resultados=rbind(resultados, data.frame(variable = var, correlacion = valor, tipo = "Numérica", mes
  }
}


for (var in cat_vars) {
  x=as.factor(df_sel$abandono)
  y=as.factor(df_sel[[var]])

  df_temp=data.frame(x = x, y = y) %>% na.omit()

  if (nlevels(df_temp$x) >= 2 && nlevels(df_temp$y) >= 2) {
    df_temp$y=fct_lump_min(df_temp$y, min = 2) %>% droplevels()
    v=DescTools::CramerV(table(df_temp$x, df_temp$y), bias.correct = TRUE)
    mes=extraer_mes(var)
    resultados=rbind(resultados, data.frame(variable = var, correlacion = v, tipo = "Categórica", mes =
  }
}


niveles_ordenados=c("7", "8", "9", "10", "11", "12", "1", "Media")


resultados$mes=factor(resultados$mes, levels = niveles_ordenados)

# Paleta de colores manual
colores_meses=c(
  "7" = "#F8766D", "8" = "#D89000", "9" = "#A3A500",
  "10" = "#39B600", "11" = "#00BF7D", "12" = "#00BFC4",
  "1" = "#619CFF", "Media" = "#BDBDBD"
)


ggplot(resultados, aes(x = reorder(variable, correlacion), y = correlacion, fill = mes)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  theme_minimal() +
  labs(title = "Correlación de variables digitales con 'abandono'",
       x = "Variable",
       y = "Correlación (Pearson o Cramér's V)",
       fill = "Mes") +
  scale_fill_manual(values = colores_meses, drop = FALSE) +
  theme(axis.text.y = element_text(size = 6))
```

Correlación de variables digitales con 'abandono'
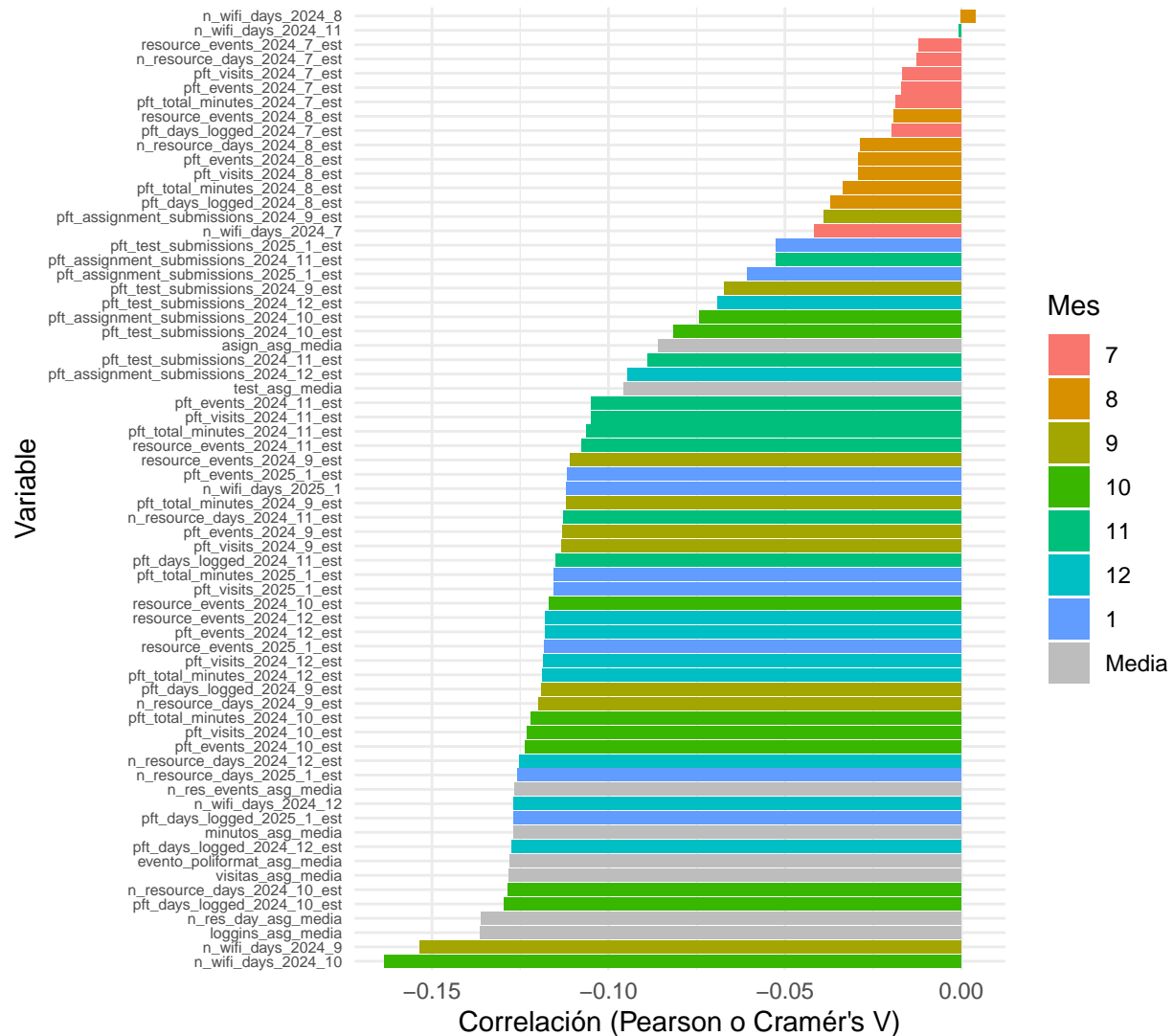
```r
lineas_multiple=poliformat_enterolineas_multiple=function(df, columnas, titulo) {

  if (!"abandono" %in% names(df)) {
    stop("La columna 'abandono' no existe en el dataframe.")
  }


  if (is.numeric(columnas)) {
    columnas=names(df)[columnas]
  }


  df$abandono=ifelse(is.na(df$abandono), 0, df$abandono)


  df_seleccionado=df[, c(columnas, "abandono"), drop = FALSE]
  df_seleccionado=df_seleccionado[sapply(df_seleccionado, is.numeric)]
```

```r
  df_valores=df_seleccionado[, setdiff(names(df_seleccionado), "abandono"), drop = FALSE]
  if (ncol(df_valores) == 0) {
    message("No hay columnas numéricas en la selección.")
    return(NULL)
  }

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

  df_melt=reshape2::melt(df_valores, id.vars = c("ID", "abandono"),
                         variable.name = "Variable", value.name = "Valor")

  df_melt$Variable=factor(df_melt$Variable, levels = columnas)

  df_negro=df_melt[df_melt$abandono == 0, ]
  df_rojo =df_melt[df_melt$abandono == 1, ]

  df_rojo=unique(df_rojo)
  df_negro=unique(df_negro)

  ggplot() +
    geom_line(data = df_negro, aes(x = Variable, y = Valor, group = ID),
              color = "black", alpha = 0.3, size = 0.5) +
    geom_line(data = df_rojo, aes(x = Variable, y = Valor, group = ID),
              color = "red", alpha = 0.8, size = 0.7) +
    labs(title =  titulo,
         x = "Variables (Columnas)", 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), "Comparación: \nEventos de Poliformat")
g2=lineas_multiple(poliformat, c(9:15), "Comparación: \nDías con Sesiones iniciadas")
g3=lineas_multiple(poliformat, c(16:22), "Comparación: \nVisitas a asignaturas")
g4=lineas_multiple(poliformat, c(23:29), "Comparación: \nMinutos Totales")
g5=lineas_multiple(poliformat, c(37:43), "Comparación: \nDías accedidos a Recursos \nde una asignatura")
g6=lineas_multiple(poliformat, c(44:50), "Comparación: \nAcciones en recursos \n de una asignatura")
g7=lineas_multiple(poliformat, c(51:55), "Comparación: \nTareas entregadas en \nuna asignatura")
g8=lineas_multiple(poliformat, c(56:60), "Comparación: \nTests entregados en \nuna asignatura")


grid_1=grid.arrange(g1, g2, ncol = 2)
```
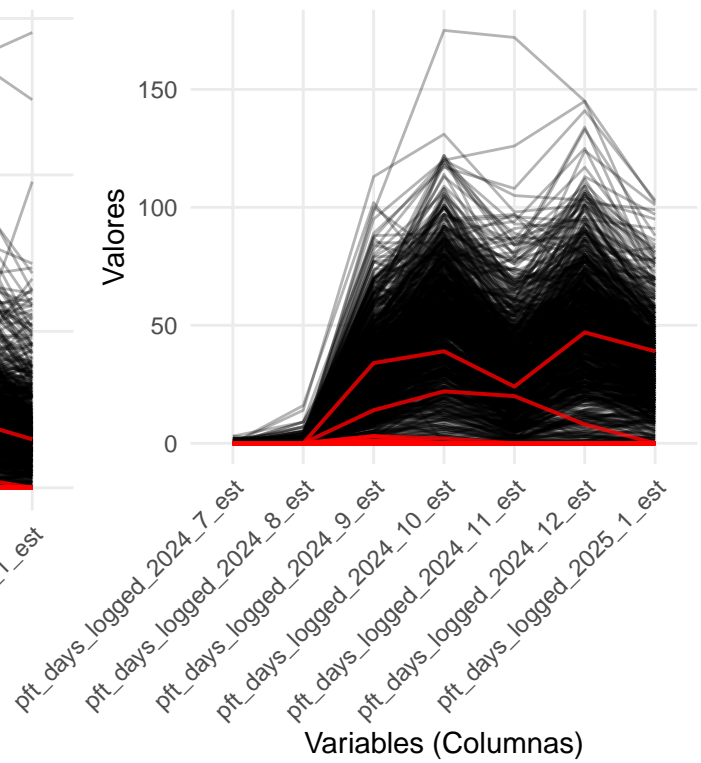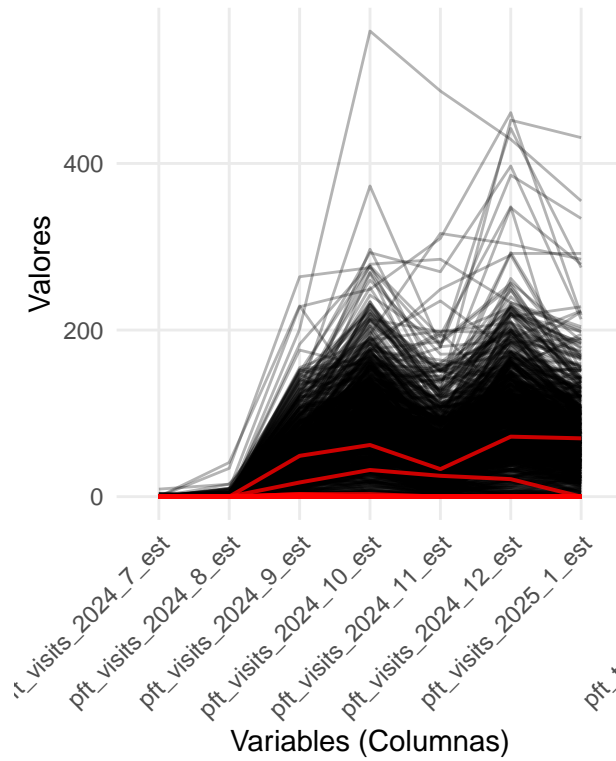
Comparación: Eventos de Poliformat

Valores / Variables (Columnas)

x-axis labels: t_events_2024_7_est, pft_events_2024_8_est, pft_events_2024_9_est, pft_events_2024_10_est, pft_events_2024_11_est, pft_events_2024_12_est, pft_events_2025_1_est
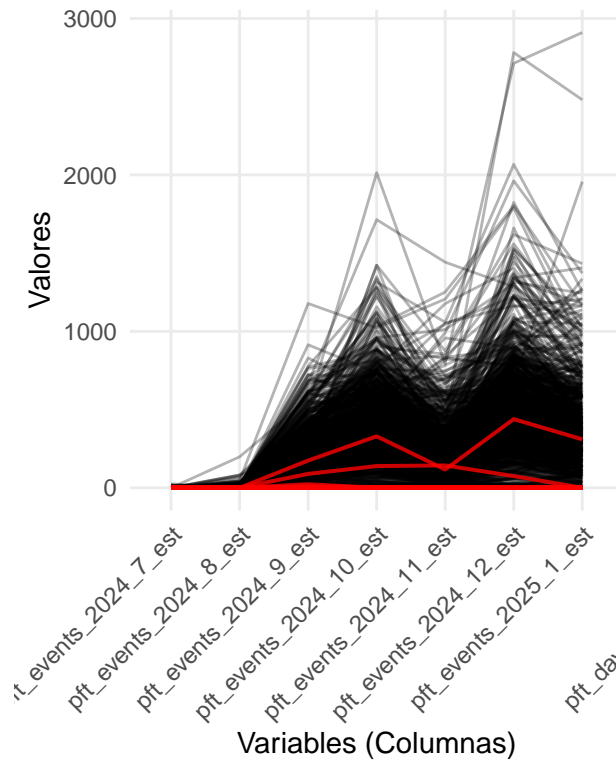
Comparación: Días con Sesiones iniciadas

Valores / Variables (Columnas)

x-axis labels: pft_days_logged_2024_7_est, pft_days_logged_2024_8_est, pft_days_logged_2024_9_est, pft_days_logged_2024_10_est, pft_days_logged_2024_11_est, pft_days_logged_2024_12_est, pft_days_logged_2025_1_est

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

Comparación: Visitas a asignaturas — Valores vs Variables (Columnas): ,t_visits_2024_7_est, pft_visits_2024_8_est, pft_visits_2024_9_est, pft_visits_2024_10_est, pft_visits_2024_11_est, pft_visits_2024_12_est, pft_visits_2025_1_est

Comparación: Minutos Totales — Valores vs Variables (Columnas): pft_total_minutes_2024_7_est, pft_total_minutes_2024_8_est, pft_total_minutes_2024_9_est, pft_total_minutes_2024_10_est, pft_total_minutes_2024_11_est, pft_total_minutes_2024_12_est, pft_total_minutes_2025_1_est
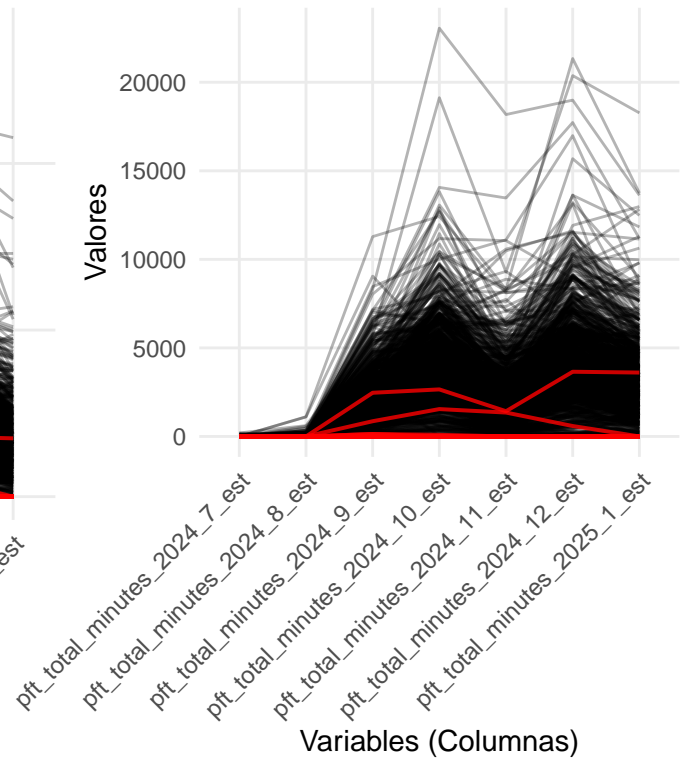
```
grid_3=grid.arrange(g5, g6, ncol = 2)
```

Comparación: Días accedidos a Recursos de una asignatura
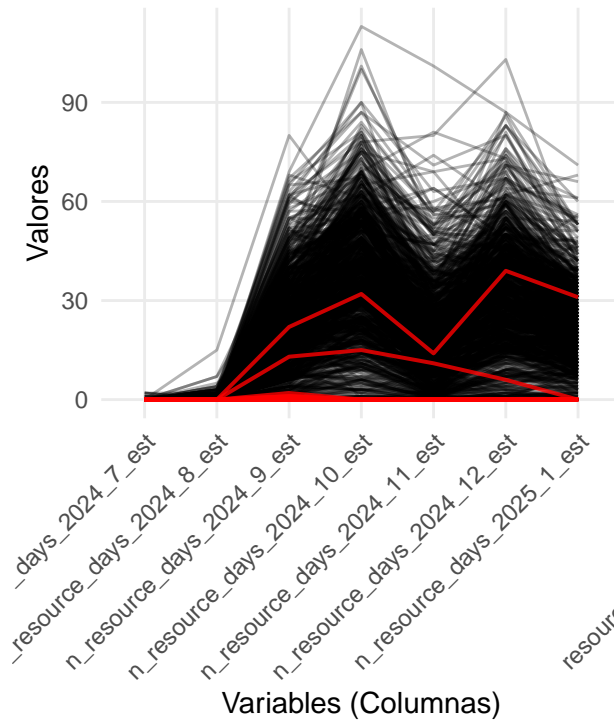
Comparación: Acciones en recursos de una asignatura
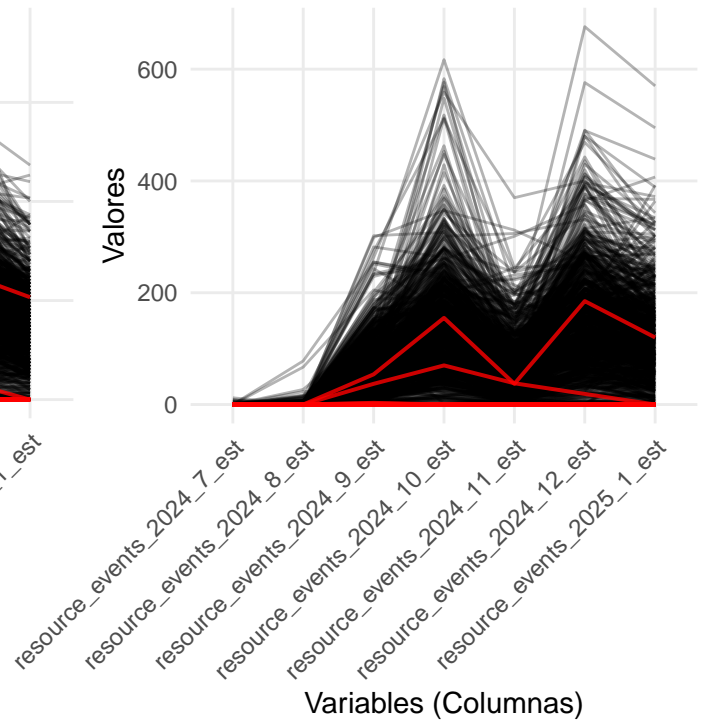
```
grid_4=grid.arrange(g7, g8, ncol = 2)
```

**Comparación:
Tareas entregadas en
una asignatura**

Valores

Variables (Columnas)

**Comparación:
Tests entregados en
una asignatura**

Valores

Variables (Columnas)
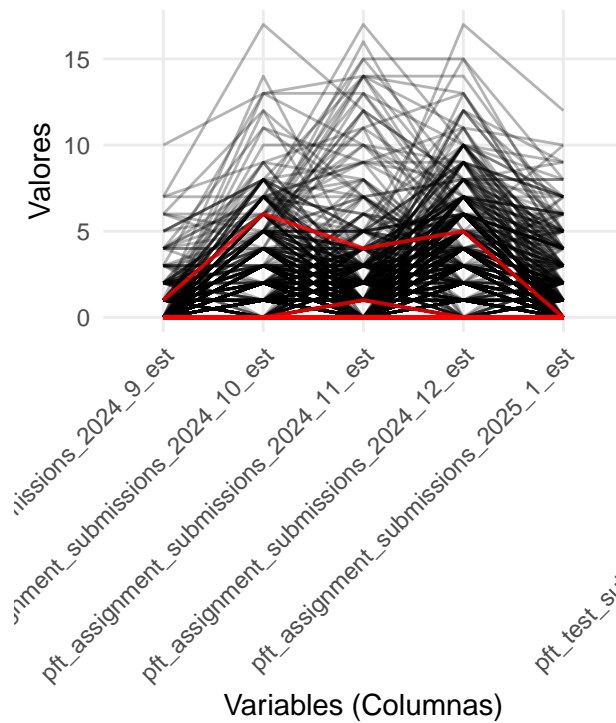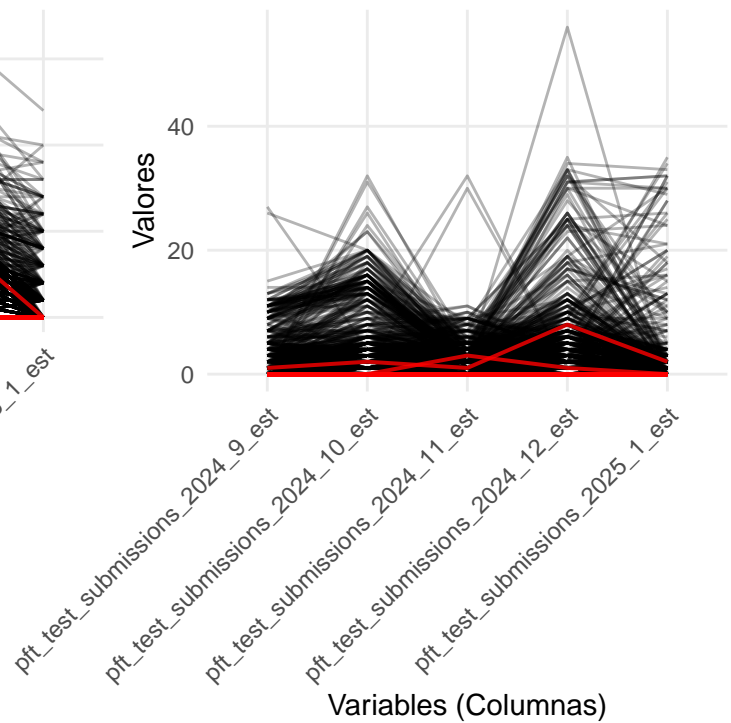
```r
lineas_multiple(poliformat, c(30:36), "Comparación: \nConexiones al wifi")
```

Comparación:
Conexiones al wifi
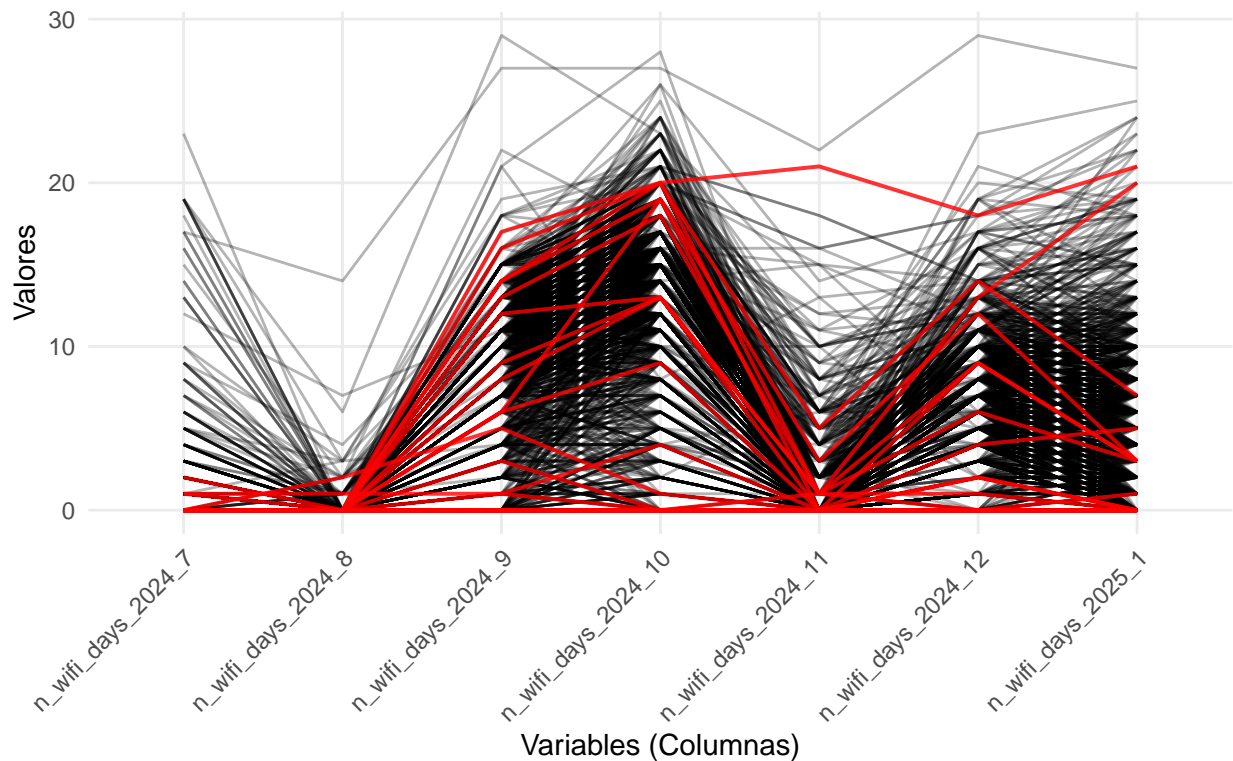

```r
variables_asg=function(df, variable, titulo, filename) {

  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, ]


  c=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))
  ggsave(paste0("./Comportamiento digital y su relación con el abandono/",filename))
  c

}

p1=variables_asg(poliformat[poliformat$abandono==1,], "evento_poliformat_asg_media", "Eventos en una asi
```

```
## Saving 10 x 10 in image
```

```
p2=variables_asg(poliformat[poliformat$abandono==1,], "loggins_asg_media", "Días con accesos a una asig
```

Saving 10 x 10 in image

```
p3=variables_asg(poliformat[poliformat$abandono==1,], "visitas_asg_media", "Visitas a una asignatura: \n
```

Saving 10 x 10 in image

```
p4=variables_asg(poliformat[poliformat$abandono==1,], "minutos_asg_media", "Minutos dedicados a una asig
```

Saving 10 x 10 in image

```
p5=variables_asg(poliformat[poliformat$abandono==1,], "n_res_day_asg_media", "Días accedidos a recursos"
```

Saving 10 x 10 in image

```
p6=variables_asg(poliformat[poliformat$abandono==1,], "n_res_events_asg_media", "Total de acciones en Re
```

Saving 10 x 10 in image

```
p7=variables_asg(poliformat[poliformat$abandono==1,], "asign_asg_media", "Tareas entregadas\nen una asig
```

Saving 10 x 10 in image

```
p8=variables_asg(poliformat[poliformat$abandono==1,], "test_asg_media", "Tests entregados \nen una asig
```

Saving 10 x 10 in image
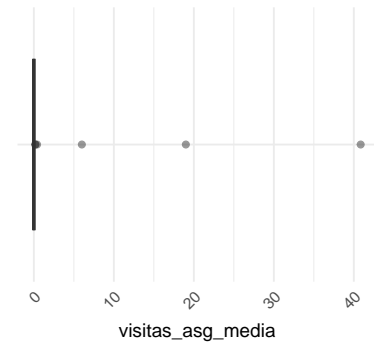
```
(p1 + p2 + p3) /(p4+p5 + p6)/(p7+p8)
```

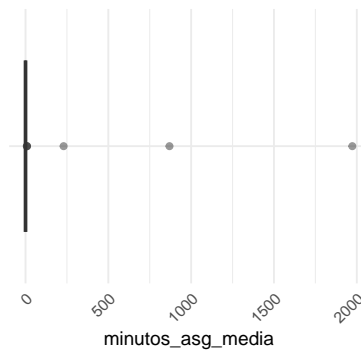Eventos en una asignatura: comparación
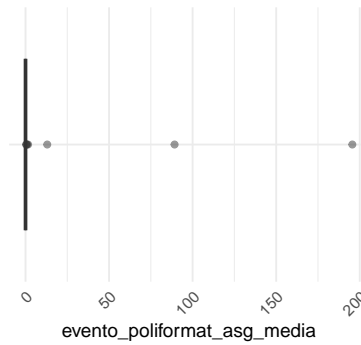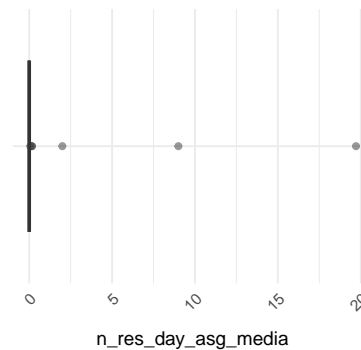
Días con accesos a una asignatura: comparación
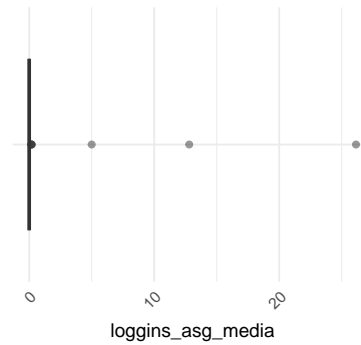
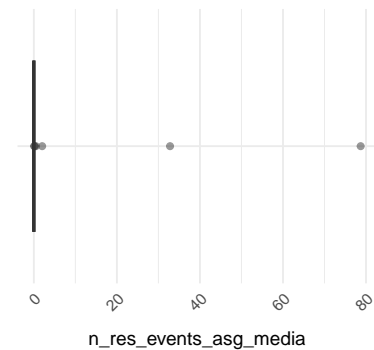Visitas a una asignatura: comparación
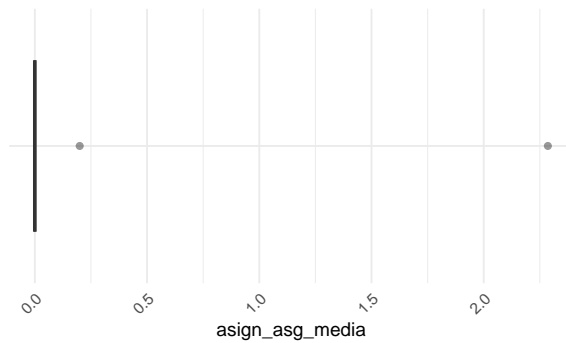
Minutos dedicados a una asignatura: comparación

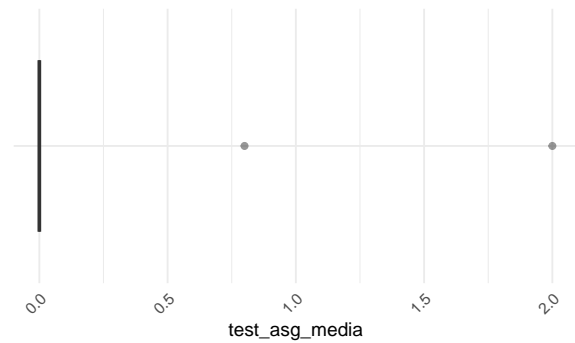Días accedidos a recursos de una asignatura: comparación

Total de acciones en Recursos de una asignatura: comparación

Tareas entregadas en una asignatura: comparación

Tests entregados en una asignatura: comparación

```
df_model=poliformat[, 2:69]



dist_mat=dist(df_model %>% select(-abandono), method = "euclidean")



a=adonis2(dist_mat ~ abandono, data = df_model, permutations = 999)
a
```

```
## Permutation test for adonis under reduced model
## Permutation: free
## Number of permutations: 999
##
```

```
## adonis2(formula = dist_mat ~ abandono, data = df_model, permutations = 999)
##           Df    SumOfSqs      R2       F Pr(>F)
## Model       1 7.4285e+08 0.01398 24.838  0.001 ***
## Residual 1752 5.2399e+10 0.98602
## Total    1753 5.3142e+10 1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
save(abandono, academicas, poliformat, sociodemografia, file="../../Datos/Capítulos/Caracterización.Rda
```