Visualización de datos

Curso básico sobre R para Investigación Social

Introducción

Luego de haber visto como manipular los datos disponibles en la base de datos del PISAC, en esta etapa del proceso del análisis de datos nos centraremos en la visualización, generando graficos de calidad utilizando el paquete `ggplot2` que forma parte del conjunto de paquetes que integran el tidyverse. Conoceremos la estructura de la sintaxis de este paquete y los argumentos necesarios para obtener los graficos que deseamos.

Instalación

Si ya tinstalamos previamente el paquete tidyverse no es necesario volver a realizar otra instalación, de lo contrario debe instalar el paquete ggplot2:

```
#Instalar paquete ggplot2
install.packages("ggplot2")
```

```
#y abrimos la librería
library(ggplot2)
#o abrimos el tidyverse
library(tidyverse)
```

La gramática de los gráficos (ggplot2)

La estructura básica de la función ggplot se compone por tres argumentos necesarios: * `data` en el cuál indicamos el dataframe del cual vamos a seleccionar las variables para graficar; * `aes` dónde indicamos el mapeo estético, es decir, qué variables vamos a graficar y en qué ejes, y * `geom` hace referencia a las capas geométricas, que a su vez, va a depender del tipo de gráfico que deseemos generar. Ejemplo:

```
ggplot(data= dataframe, aes(x= variable1, y= variable2)) + geom_point
```

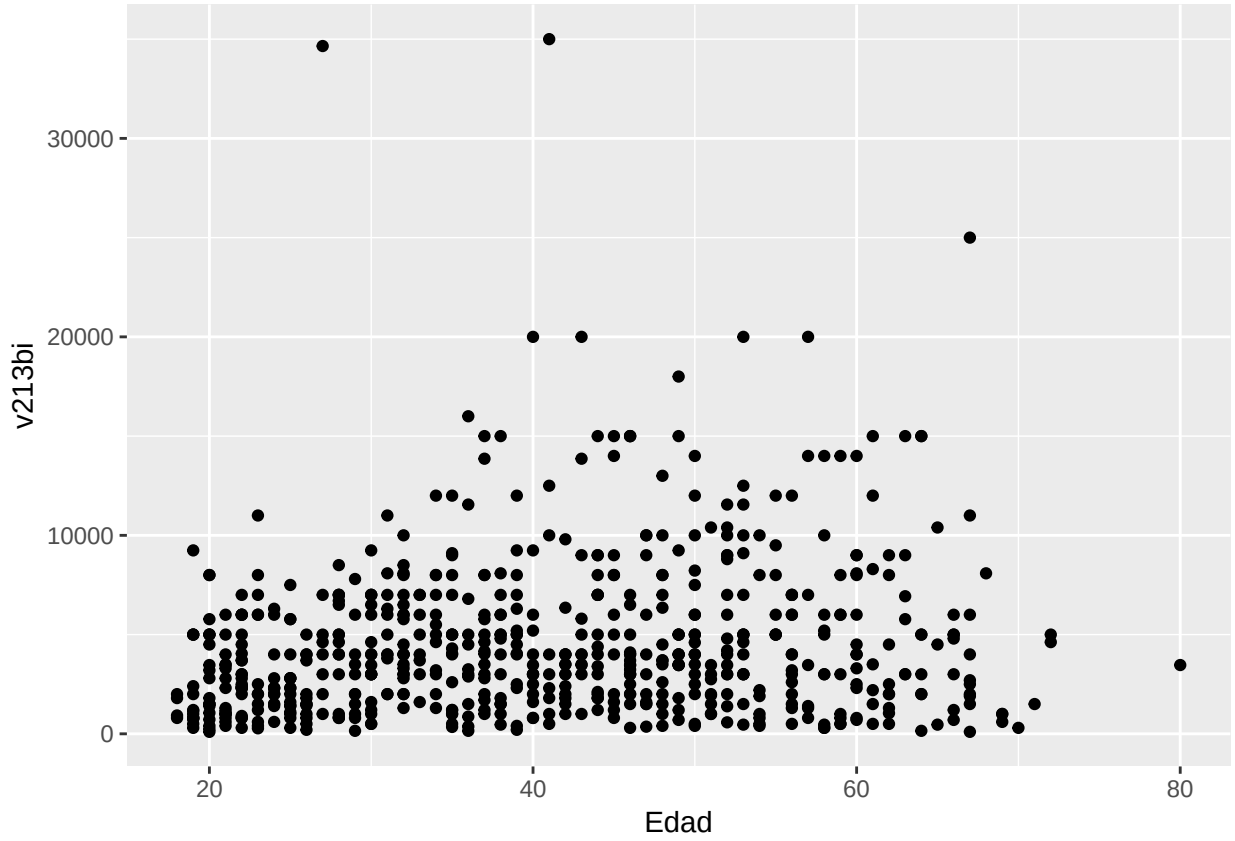
Gráfico de dispersión

Hagamos la prueba con un gráfico de dispersión utilizando las variables Ingreso neto de la Ocupación Principal (v213bi) y la Edad. Pero, previamente vamos a crear un nuevo dataframe que va a contener solamente las variables que necesitamos en el gráfico. Para ello usaremos el paquete `dplyr`.

```
#Abrimos la librería
library(dplyr)
data.graf1 <- NEA %>% filter(v213bi > 0 &              #filtramos solo los ingr. > a 0
                             Edad >= 18 &              #y las edades > o = a 18 años
                             !is.na(Clase_recod)) %>% #omitimos los casos NA en Clase_recod
  select(v213bi, Edad, Sexo, Clase_recod) %>%          #seleccionamos otras variables que
                                                       #necesitaremos para graficar
  rename(Clase = Clase_recod)                          #cambiamos el nombre de la variable
```

Entonces, a partir de este dataframe vamos a realizar el grafico:

```
ggplot(data= data.graf1, aes(x= Edad, y= v213bi)) + geom_point()
```



En un principio la estética del gráfico es bastante simple, pero podemos ir mejorando a nuestro gusto agregando más parámetros a la sintaxís de la función, utilizando siempre el símbolo +.

Tabla 6

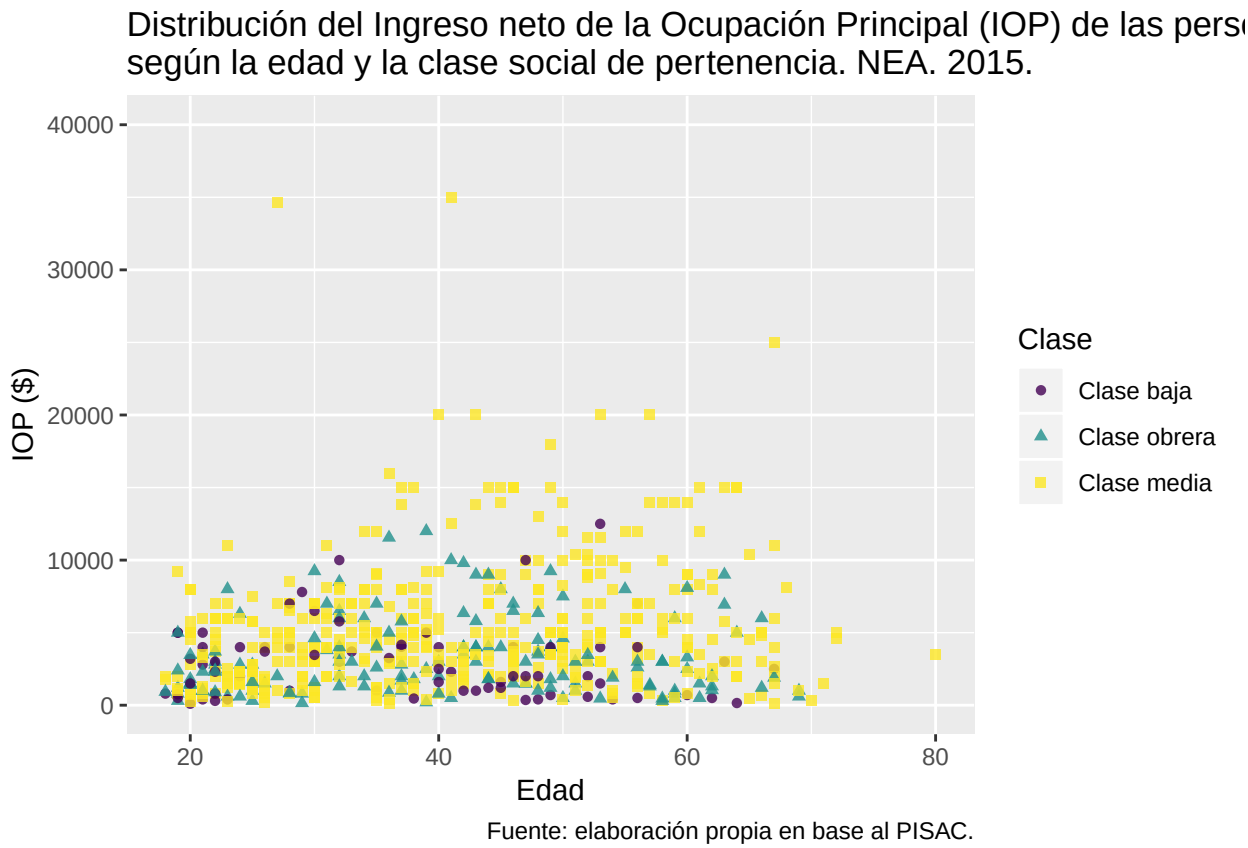
Parámetro	Resultado
labs()	Etiquetas: title= "titulo", x= "etiqueta eje x", y= "etiqueta eje y", caption= "texto al pie"
theme()	tema del gráfico: base_family= "fuente", base_size= tamaño fuente
scale_x_continuous()	Ajustar eje x
scale_y_continuous()	Ajustar eje y
geom_point()	Gráfico de dispersión
geom_bar()	Gráfico de barras
geom_histogram()	Histograma
geom_boxplot()	Diagrama de caja
geom_freqpoly()	Polígono de frecuencia

En la *Tabla 6* se pueden ver sólo algunos de los parámetros de la función ggplot. Si desea apreciar el resto de

las propiedades estéticas de los gráficos, puede consultar la *hoja de referencia* del paquete aquí. Retomemos el gráfico anterior para mejorar su estética

```
Grafico.1 <- ggplot(data= data.graf1, aes(x= Edad, y= v213bi)) +
  geom_point(aes(color= Clase,              #agregamos una tercer variable
                 shape= Clase),             #formas geometricas de la nube
                 alpha= 0.8) +              #transparencia
  scale_y_continuous(limits = c(0,40000)) + #ajustamos limites del eje y
  labs(title = "Distribución del Ingreso neto de la Ocupación Principal (IOP) de las personas
según la edad y la clase social de pertenencia. NEA. 2015.",
       y= "IOP ($)",
       caption = "Fuente: elaboración propia en base al PISAC.") #agregamos etiquetas

Grafico.1
```

```
## Warning: Using shapes for an ordinal variable is not advised
```

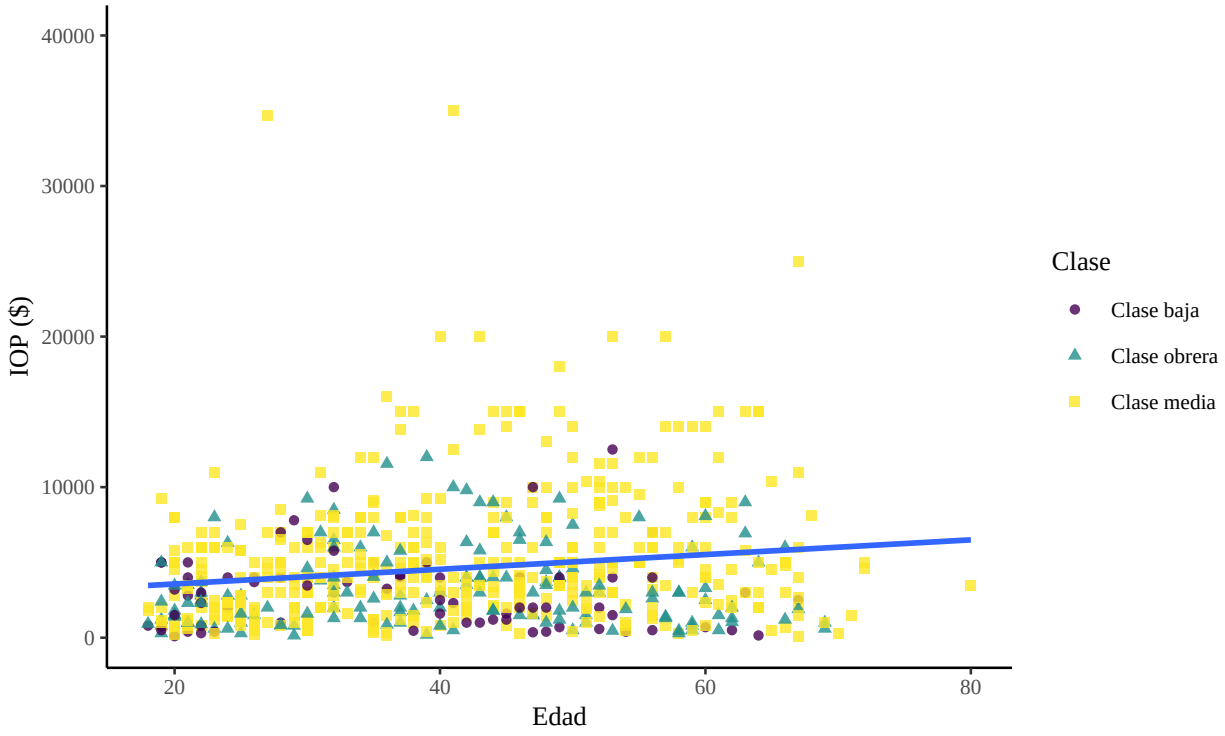


Observe que además de agregar argumentos a la función, guardamos el gráfico en un objeto llamado *Grafico.1*. Ahora podemos seguir agregando argumentos a partir de este objeto.

```
Grafico.1 <- Grafico.1 + geom_smooth(method = lm, se = FALSE) +  #incorporamos la recta de ajuste
  theme_classic(base_family = "serif",     # agregamos un tema y modificamos la fuente
                base_size = 10)
Grafico.1
```

```
## Warning: Using shapes for an ordinal variable is not advised
```

Distribución del Ingreso neto de la Ocupación Principal (IOP) de las personas según la edad y la clase social de pertenencia. NEA. 2015.



Fuente: elaboración propia en base al PISAC.

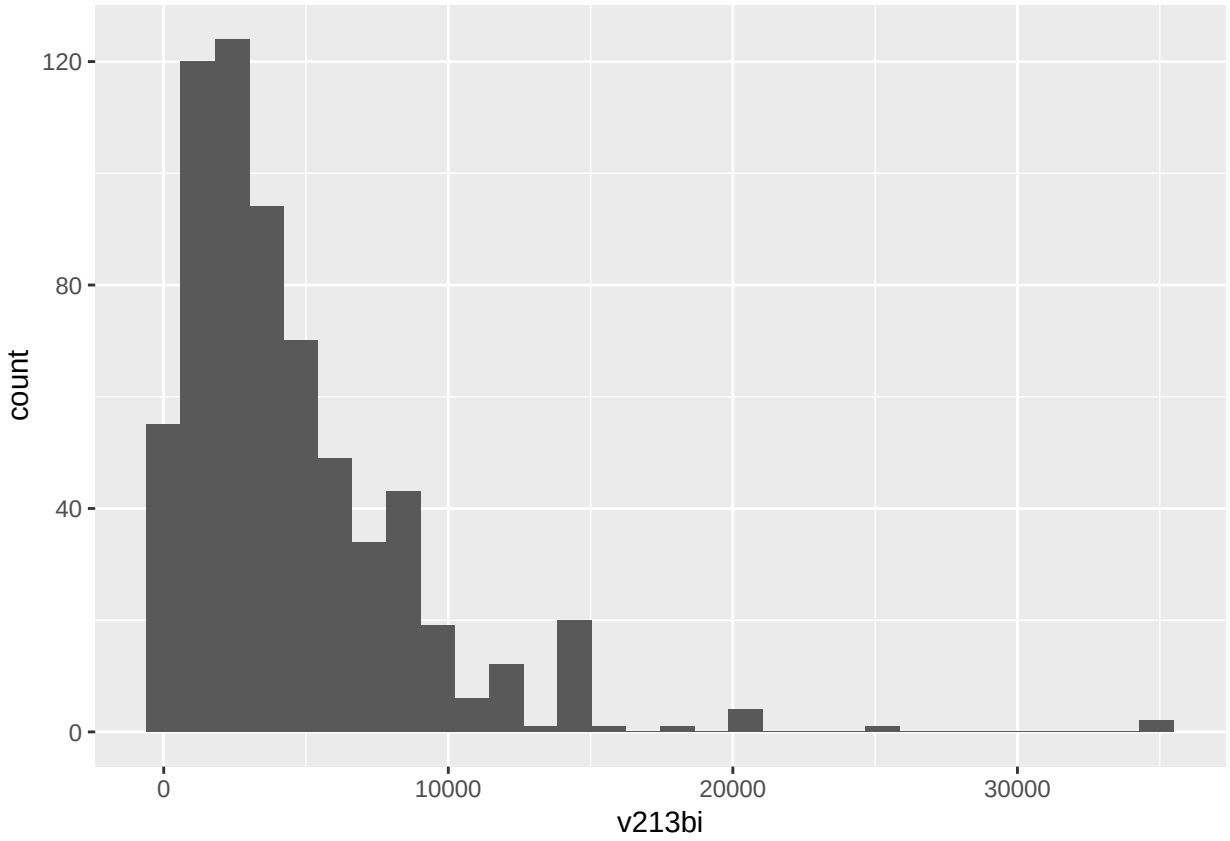
Histograma

Vamos a crear un histograma utilizando la variable Ingreso neto de la Ocupación Principal, pero primeramente creamos el dataframe que necesitamos para graficar.

```
data.graf2 <- NEA %>% filter(v213bi > 0 &
                             !is.na(Clase_recod)) %>%
  select(v213bi, Edad_recod, Clase_recod, f_calib3.x) %>%
  rename(Edad = Edad_recod)
```

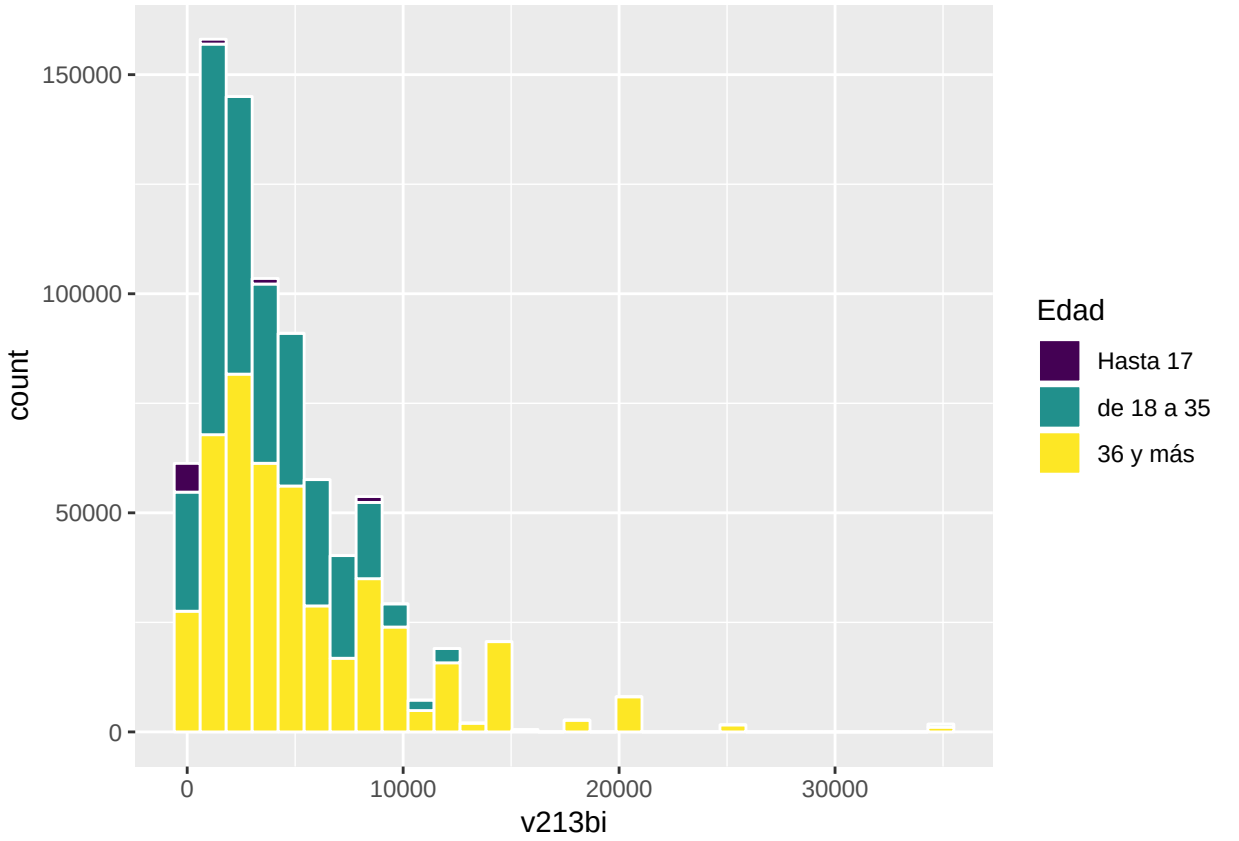
Y graficamos en este caso usando `geom_histogram`

```
ggplot(data= data.graf2, aes(x= v213bi)) + geom_histogram()
```



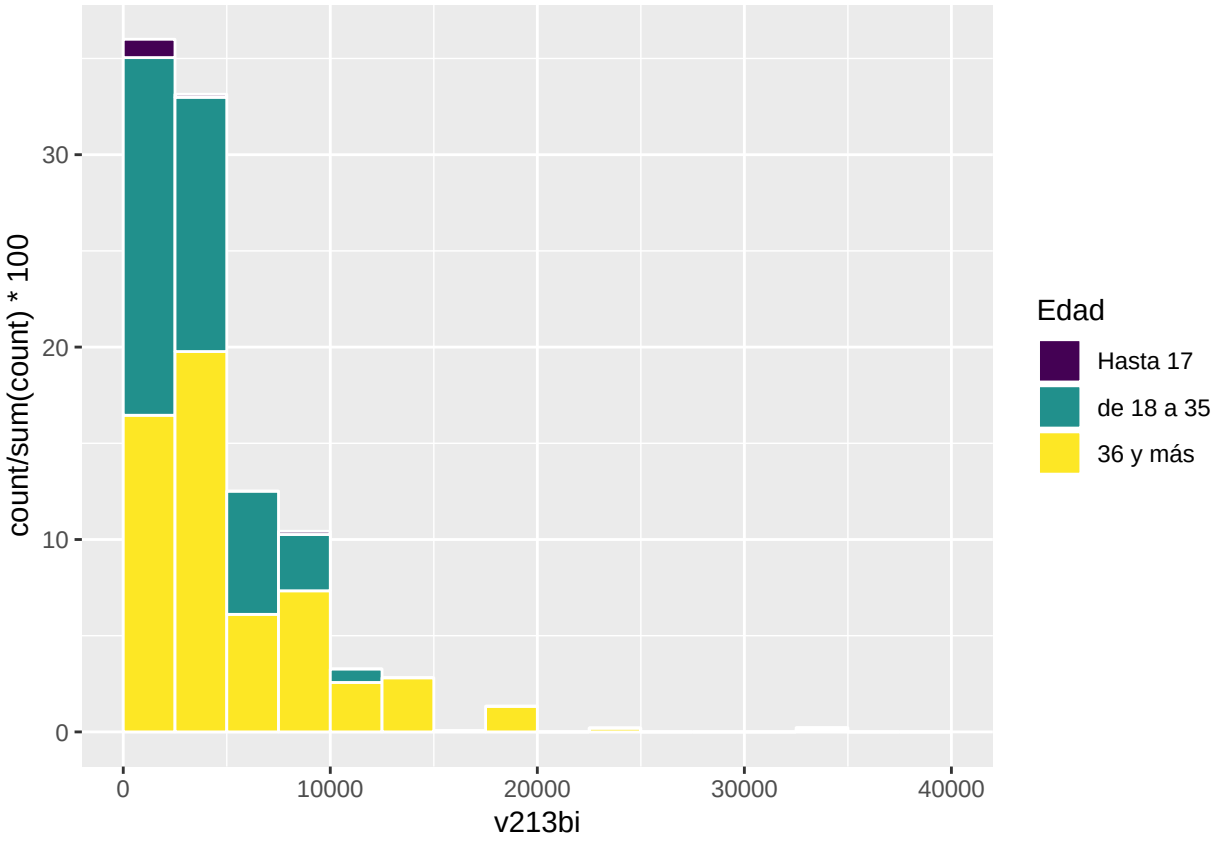
Podemos expandir los casos usando el argumento `weight`. Para esta base de datos, la variable ponderadora de casos es *factor de calibración* (f_calib3); por lo tanto, la incorporamos a la función y agregamos una segunda variable que es la Edad.

```
ggplot(data= data.graf2, aes(x= v213bi, weight= f_calib3.x)) +
  geom_histogram(aes(fill = Edad),
                 color = "white") #las lineas en color blanco
```



Observe que ahora las frecuencias de las barras han aumentado porque se expandieron los casos, además las barras se dividen en función a los valores de la tercer variable Edad. También, como se habrá podido dar cuenta, las frecuencias de las barras del histograma están expresadas en valores absolutos. Si lo que deseamos es que las frecuencias se expresen valores porcentuales, necesitamos modificar la escala del del eje y.

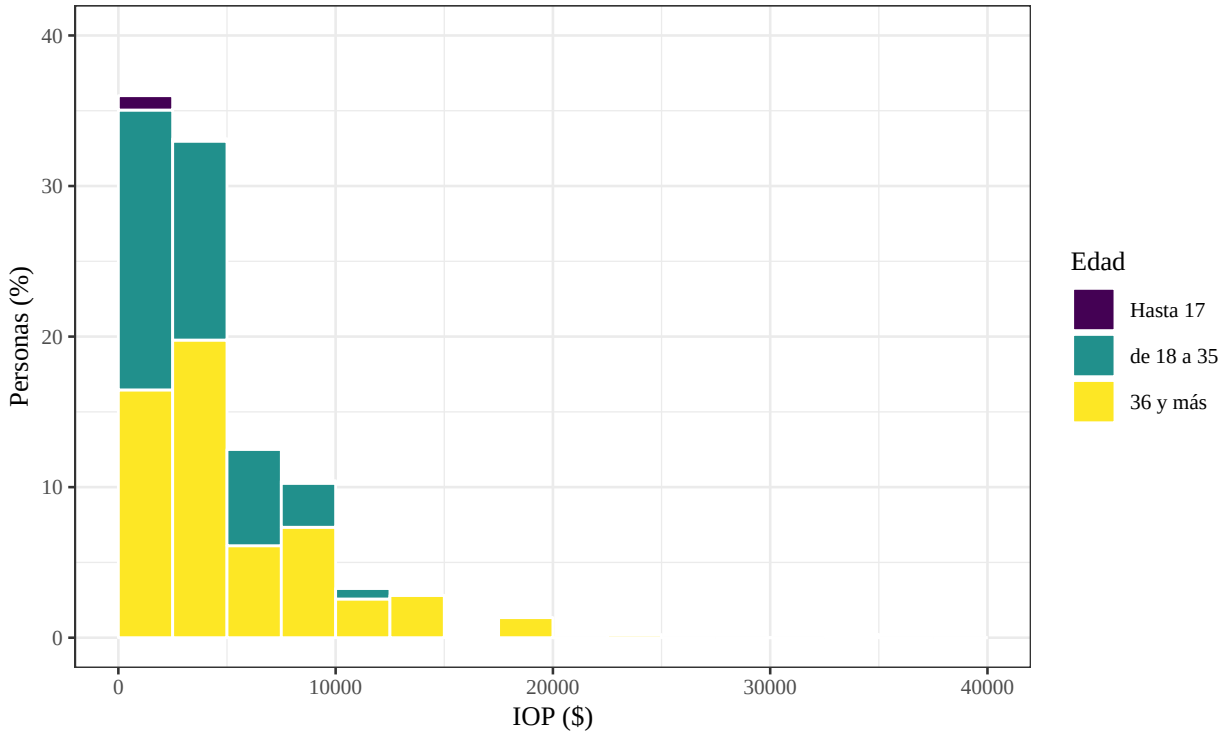
```
#Abrimos la librería
library(scales)

#Ahora procedemos a modificar las frecuencias
Grafico.2 <- ggplot(data= data.graf2, aes(x= v213bi, weight= f_calib3.x)) +
  geom_histogram(aes(fill = Edad,
                     y= stat(..count..)/sum(..count..)*100), #cambiamos eje a %
                 color = "white",
                 breaks = seq(0,40000, by= 2500)) #definimos amplitud de intervalos
Grafico.2
```



Modificamos los limites de los ejes, agregamos las etiquetas.

```
Grafico.2 <- Grafico.2 +
  scale_x_continuous(limits = c(0,40000)) +
  scale_y_continuous(limits = c(0,40)) +
  labs(title = "Distribución del Ingreso de la Ocupación Principal de las personas según
la edad. NEA. 2015.",
       x= "IOP ($)",
       y= "Personas (%)",
       caption = "Fuente: elaboración propia en base a PISAC") +
  theme_bw(base_family = "serif",
           base_size = 10)
Grafico.2
```

Distribución del Ingreso de la Ocupación Principal de las personas según la edad. NEA. 2015.



Fuente: elaboración propia en base a PISAC

Grafico de barras

Hagamos un grafico de barras con la variable *Nivel educativo* pero primero vamos a recategorizar los valores:

```
#Primero consultamos los valores
levels(NEA$nivel_ed)
```

```
##  [1] "Menores de 5 años"
##  [2] "Sin instrucción (incluye nunca asistió o sólo asistió a sala de 5)"
##  [3] "Primaria/EGB incompleto"
##  [4] "Primaria/EGB completo"
##  [5] "Secundario/Polimodal incompleto"
##  [6] "Secundario/Polimodal completo"
##  [7] "Terciario incompleto"
##  [8] "Terciario completo"
##  [9] "Universitario incompleto"
## [10] "Universitario completo"
## [11] "Educación especial"
## [12] "NS/NR"
```

```
#Recategorizamos
NEA <- NEA %>% mutate(nivel_ed_recod =
                        case_when(
                          nivel_ed %in% c("Menores de 5 años",
```

```
                                        "Sin instrucción (incluye nunca asistió o sólo asistió
                                        a sala de 5)",
                                        "Primaria/EGB incompleto",
                                        "Primaria/EGB completo",
                                        "Secundario/Polimodal incompleto") ~
                            "Secundaria incompleta o menos",
                          nivel_ed == "Secundario/Polimodal completo" ~ "Secundaria completa",
                          nivel_ed %in% c("Terciario incompleto",
                                        "Universitario incompleto") ~ "Terc/Univ. incompleto",
                          nivel_ed %in% c("Terciario completo",
                                        "Universitario completo") ~ "Terc/Univ. completo",
                          nivel_ed == "Educación especial" ~ "Educación especial",
                          nivel_ed == "NS/NR" ~ "NS/NR")) %>%
  mutate(nivel_ed_recod = ordered(nivel_ed_recod,
                                  levels = c("Secundaria incompleta o menos",
                                             "Secundaria completa",
                                             "Terc/Univ. incompleto",
                                             "Terc/Univ. completo",
                                             "Educación especial",
                                             "NS/NR")))
levels(NEA$nivel_ed_recod)
```

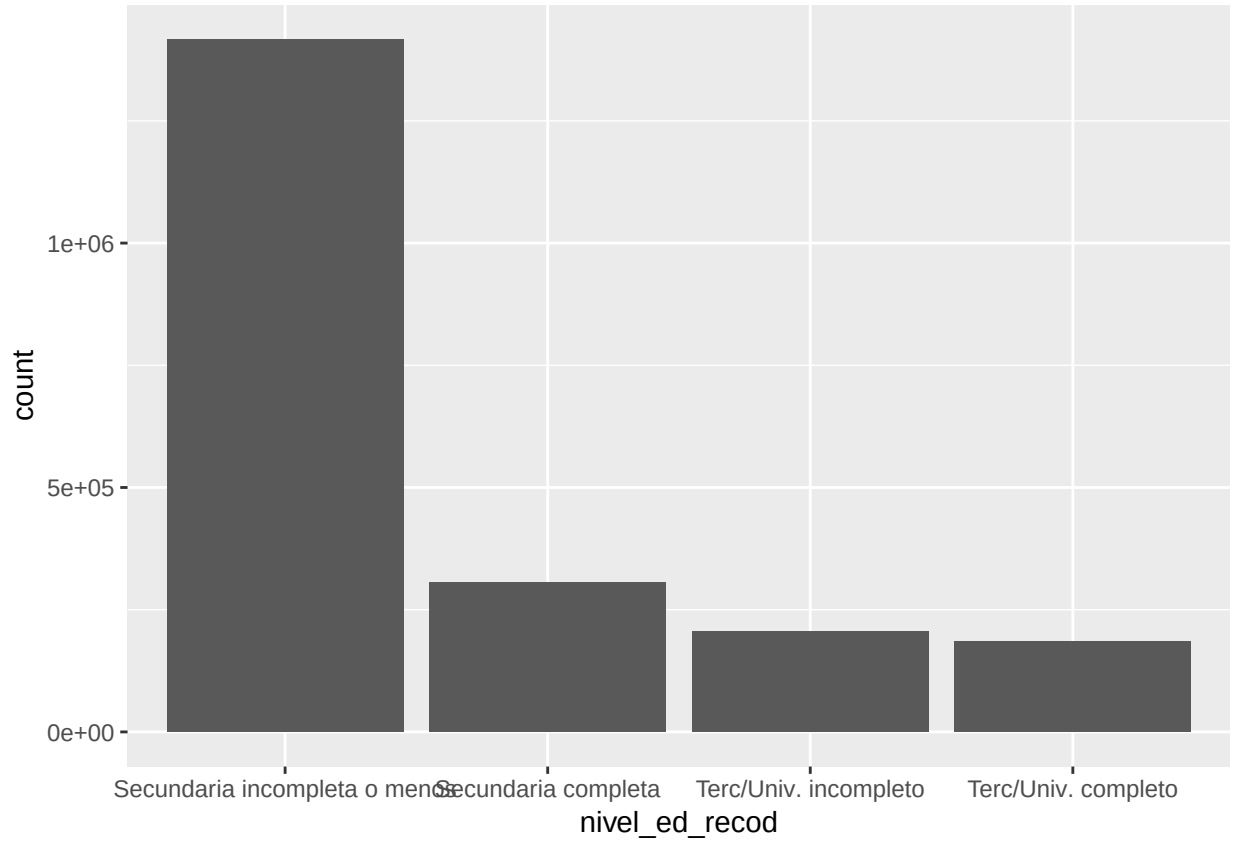
```
## [1] "Secundaria incompleta o menos" "Secundaria completa"
## [3] "Terc/Univ. incompleto"         "Terc/Univ. completo"
## [5] "Educación especial"            "NS/NR"
```

Recategorizamos los valores en seis categorías, ahora procederemos a crear el dataframe para graficar:

```
data.graf3 <- NEA %>%
  filter(nivel_ed_recod %in% c("Secundaria incompleta o menos",
                               "Secundaria completa",
                               "Terc/Univ. incompleto",
                               "Terc/Univ. completo") &
           Sexo != "Otro" &
           Clase_recod != "Clase alta" &
           !is.na(Est_civil) &
           !is.na(Clase_recod)) %>%
  select(nivel_ed_recod, Sexo, Est_civil, f_calib3.x, Clase_recod)
```

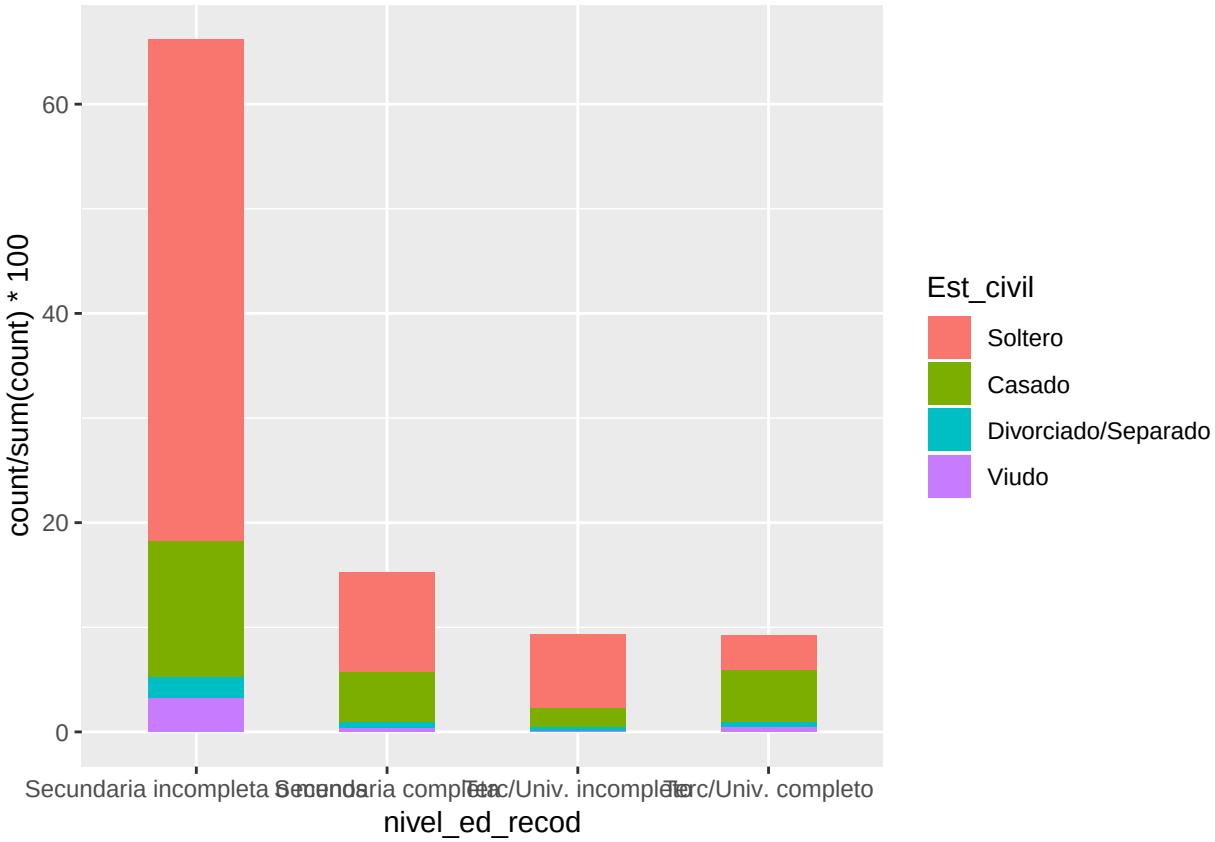
Para este gráfico usamos `geom_bar`:

```
ggplot(data.graf3, aes(x= nivel_ed_recod, weight = f_calib3.x)) +
  geom_bar()
```



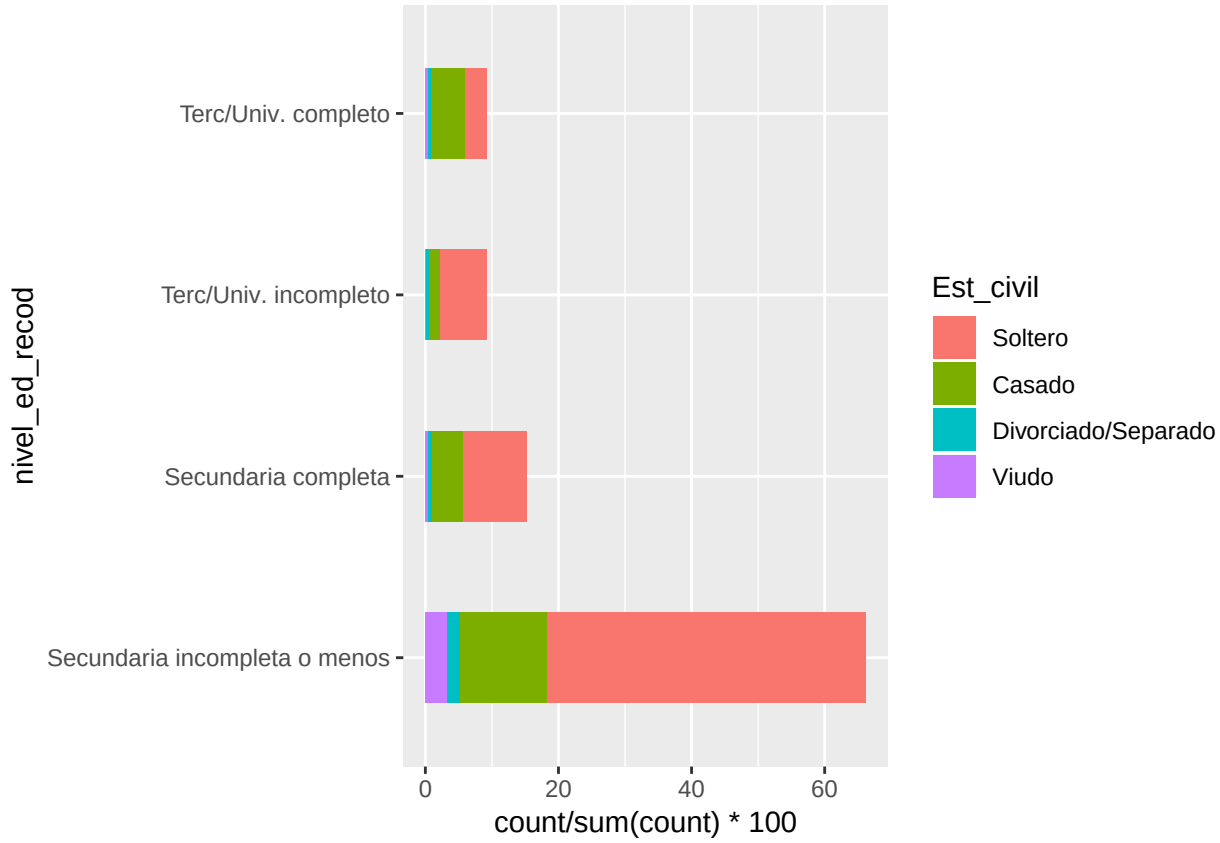
Y comenzamos a mejorarlo...

```
Grafico.3 <- ggplot(data.graf3, aes(x= nivel_ed_recod)) +
  geom_bar(aes(y = ..count../sum(..count..)*100,
               fill = Est_civil),
           width = 0.5)   #disminuyo el ancho de las barras
Grafico.3
```



Agregamos un tema y cambiamos la orientación de las barras:

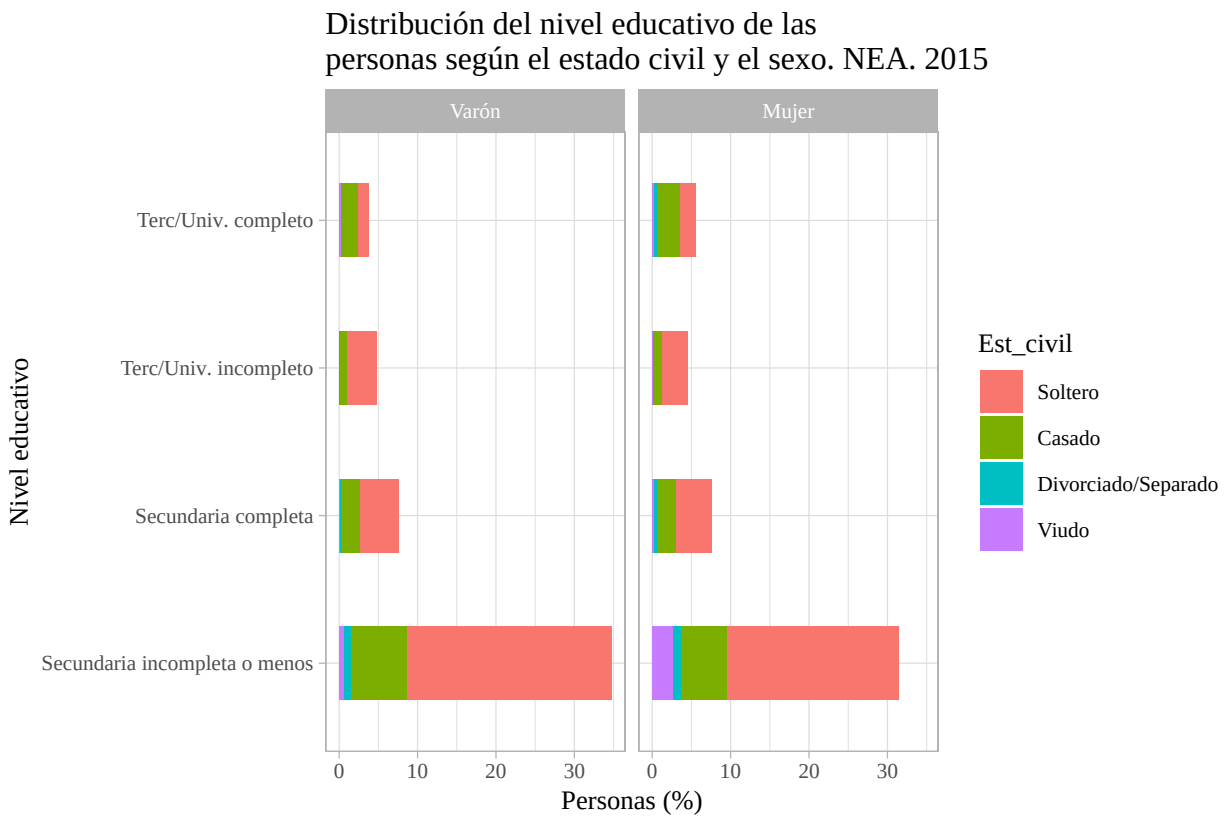
```
Grafico.3 <- Grafico.3 +
  coord_flip()  #cambio la orientación de las barras a horizontal
Grafico.3
```



Facetas

Las facetas nos permiten separar la distribucion de la variable en sub-poblaciones en función a las categorías de otra u otras variables adicionales. Hay dos formas de dividir los gráficos en facetas; `facet_wrap()`: nos permite dividir las sub-poblaciones a partir de las categorías de una variable adicional; `facet_grid()`: nos permite dividir las sub-poblaciones a partir de las categorías de dos variables adicionales. Veamos algunos ejemplos continuando con el mismo gráfico.

```
Grafico.3.a <- Grafico.3 +
  facet_wrap(~Sexo) +  #separamos los sub-grupos por las
                       #categorías de la variable Sexo
  labs(title= "Distribución del nivel educativo de las
personas según el estado civil y el sexo. NEA. 2015",
       x= "Nivel educativo",
       y= "Personas (%)",
       caption = "Fuente: elaboración propia en base a PISAC.") +
  theme_light(base_family = "serif",
              base_size = 10)
Grafico.3.a
```

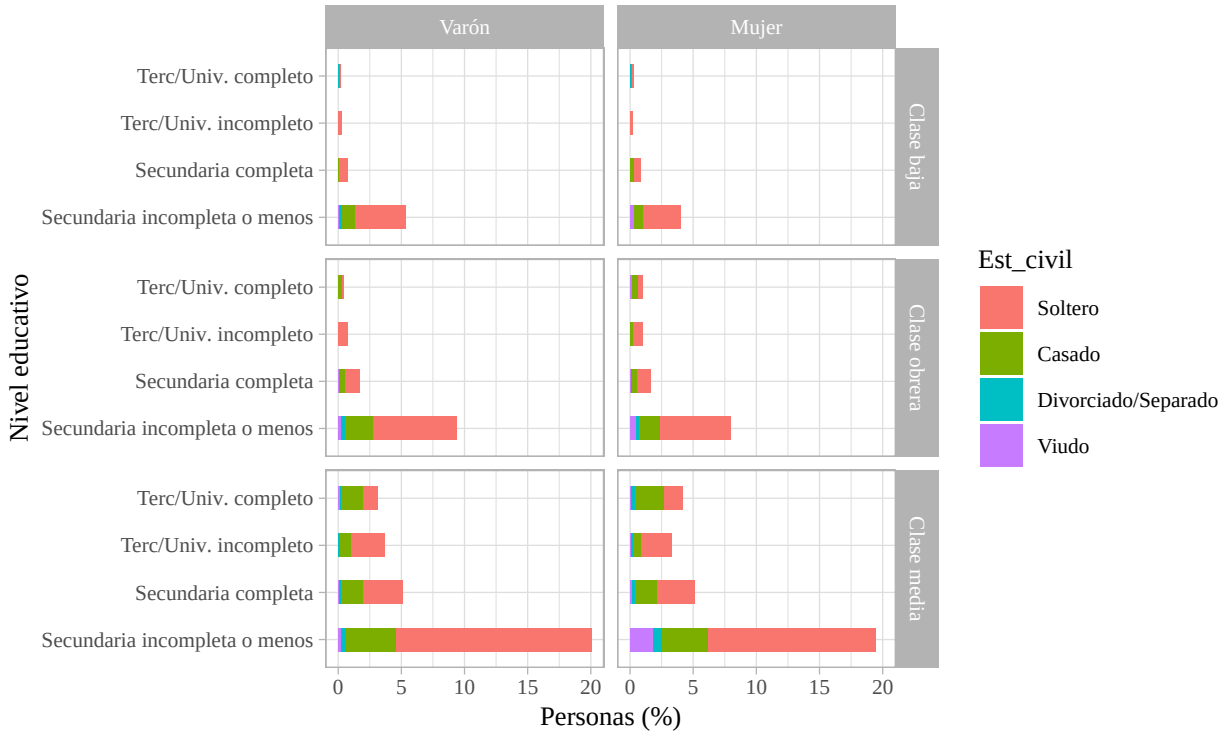


Fuente: elaboración propia en base a PISAC.

Incluyendo dos variables:

```
Grafico.3.b <- Grafico.3 +
  facet_grid(Clase_recod ~ Sexo) + #separamos los sub-grupos por las
                                   #categorías de las var. Clase_recod y Sexo
  labs(title= "Distribución del nivel educativo de las personas según el estado
civil, el sexo y la clase social de pertenencia. NEA. 2015",
       x= "Nivel educativo",
       y= "Personas (%)",
       caption = "Fuente: elaboración propia en base a PISAC.") +
  theme_light(base_family = "serif",
              base_size = 10)
Grafico.3.b
```

Distribución del nivel educativo de las personas según el estado civil, el sexo y la clase social de pertenencia. NEA. 2015



Fuente: elaboración propia en base a PISAC.

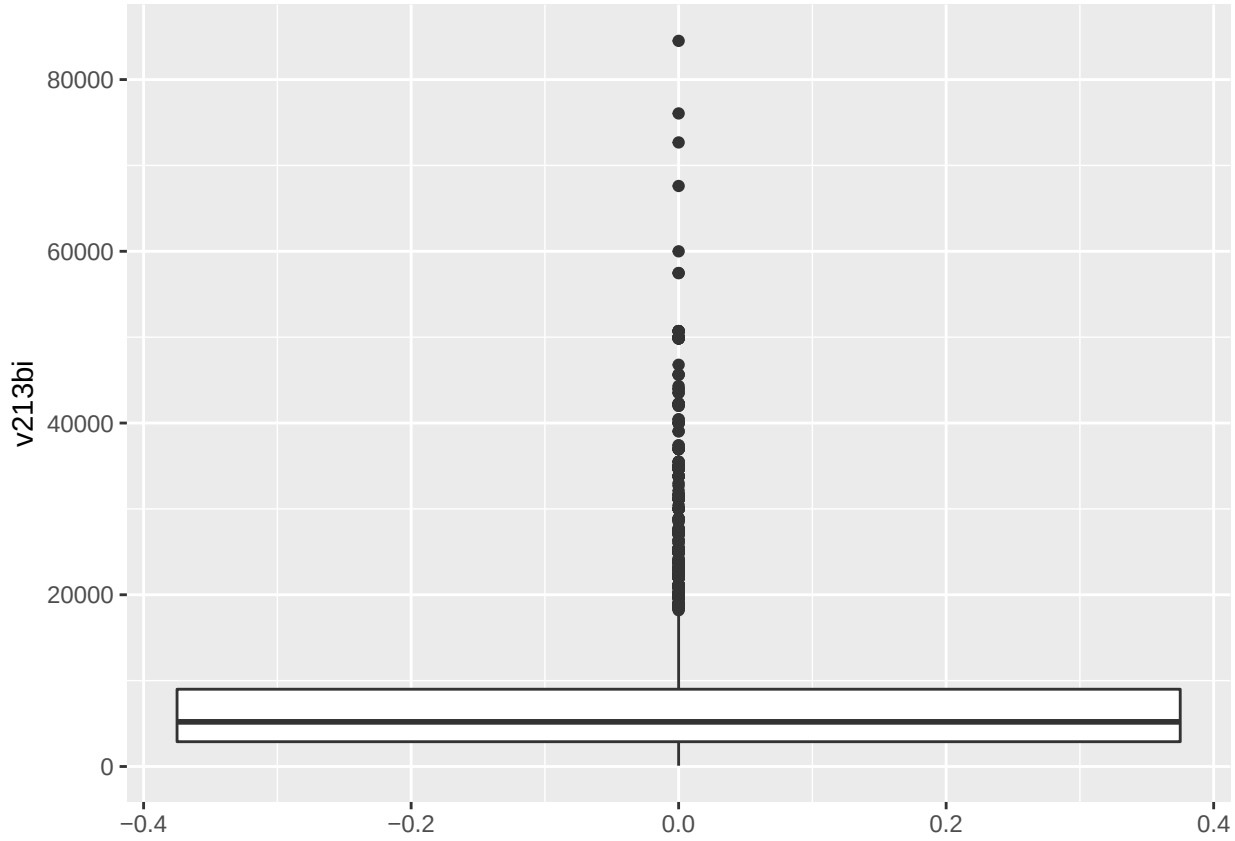
Paletas de colores

Hay una serie de paletas que podemos añadir a los gráficos para cambiar los colores, indicando la paleta que más nos agrade dentro del argumento `scale_fill_brewer`. En este enlace puede apreciar algunas paletas del paquete *RColorBrewer*. A continuación, veamos un ejemplo de cómo incorporar una paleta de color, graficando un box-plot.

Box-plot

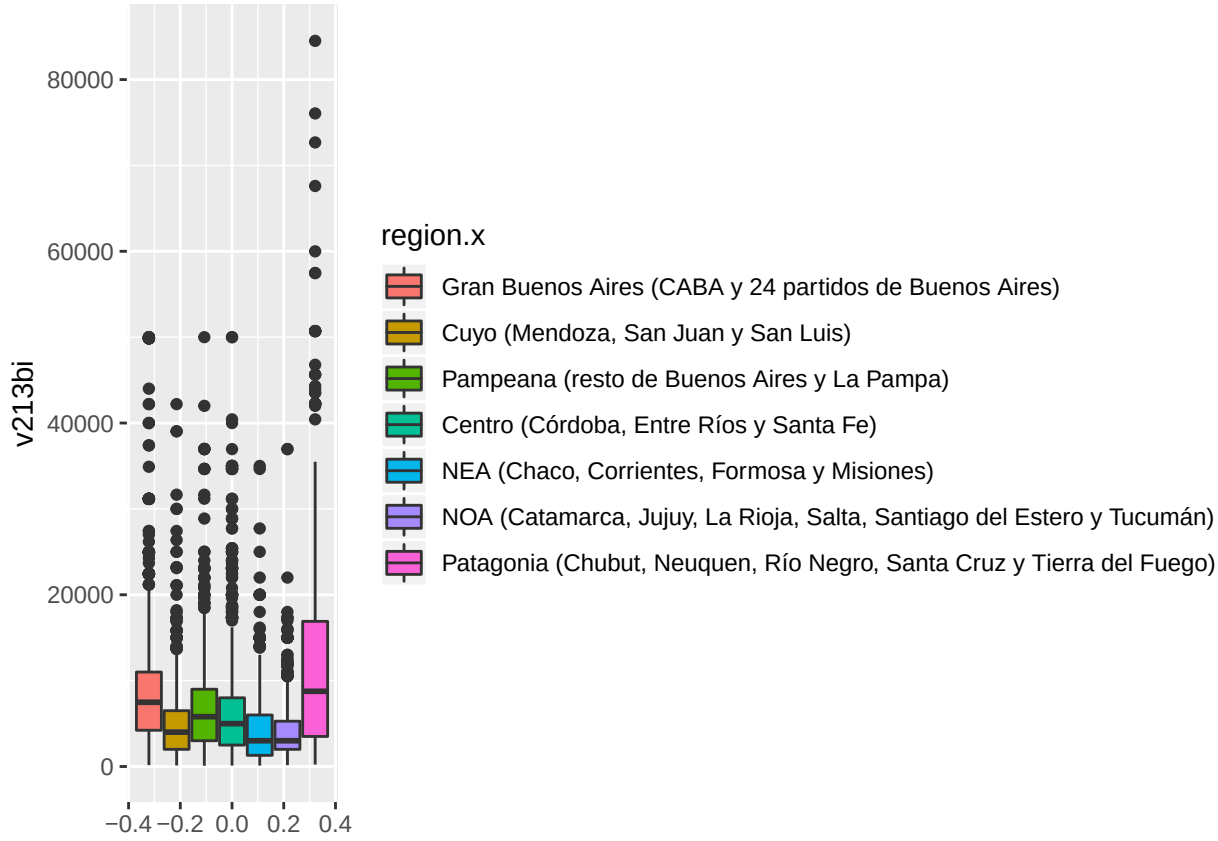
Como paso previo a realizar el grafico, vamos a crear el dataframe que necesitamos para graficar y para ello, retomemos la base de datos **pisac** que unimos al principio del trabajo:

```
data.graf4 <- pisac %>%
  filter(v213bi > 0,
         !is.na(v237)) %>%
  select(v213bi, region.x, f_calib3.x, v237)

#Graficamos
ggplot(data= data.graf4, aes(y= v213bi, weight= f_calib3.x)) + geom_boxplot()
```



Observemos la distribución del ingreso por región:

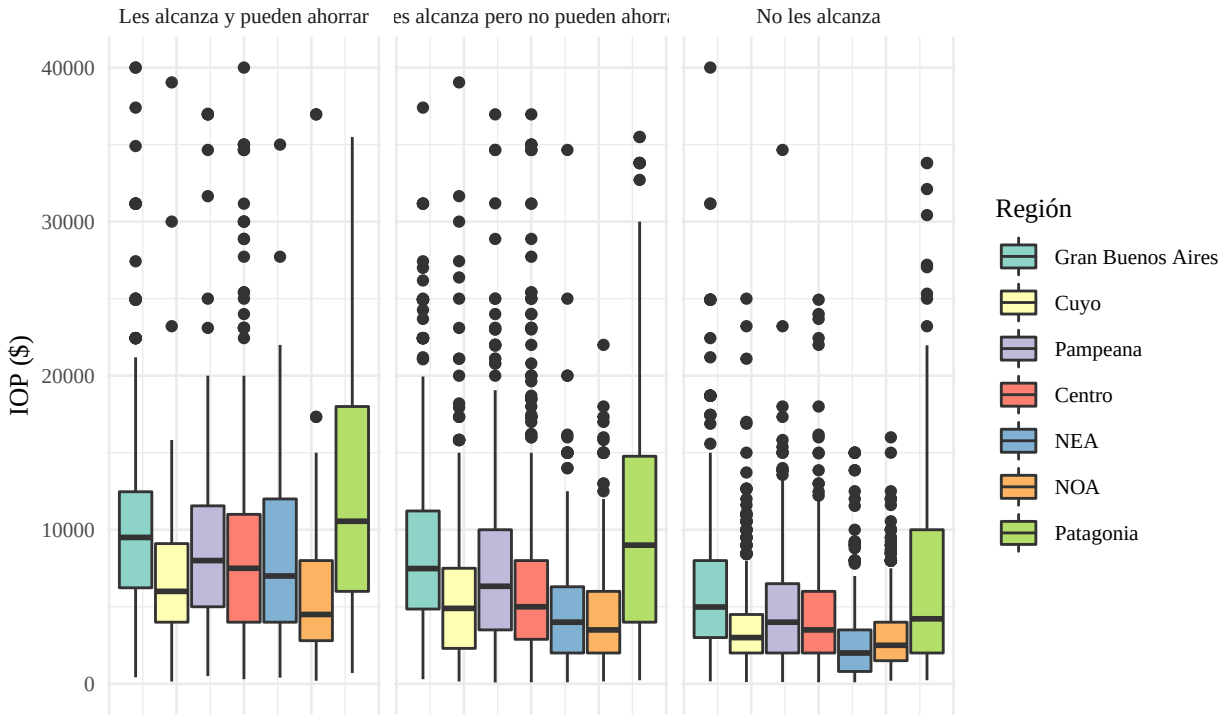
```
Grafico.4 <- ggplot(data= data.graf4, aes(y= v213bi, weight= f_calib3.x)) +
  geom_boxplot(aes(fill = region.x))
Grafico.4
```



Agregamos la paleta *Set3* y dividimos por facetas:

```
Grafico.4 <- Grafico.4 +
  scale_fill_brewer(palette = "Set3",  #modificamos los colores cambiando de paleta
                    name = "Región",
                    labels = c("Gran Buenos Aires",
                               "Cuyo",                #renombramos las etiquetas
                               "Pampeana",
                               "Centro",
                               "NEA",
                               "NOA",
                               "Patagonia")) +
  facet_wrap(~ v237) +
  scale_y_continuous(limits = c(0,40000)) +  #restringimos los ingresos hasta 40000 pesos
  theme_minimal(base_family = "serif",
                base_size = 10) +
  labs(title = "Distribución del Ingreso neto de la Ocupación Principal (IOP) de las
personas según la region y la percepción del ingreso. Argentina. 2015.",
       y = "IOP ($)",
       caption = "Fuente: elaboración propia en base a PISAC") +
  theme(axis.text.x=element_blank())  #eliminamos valores del eje x
  Grafico.4
```

Distribución del Ingreso neto de la Ocupación Principal (IOP) de las personas según la region y la percepción del ingreso. Argentina. 2015.



Fuente: elaboración propia en base a PISAC

Referencias

Este post está basado en los siguientes aportes bobliográficos:

WICKHAM H., GROLEMUND G. "R for Data Science" (2017). Recuperado de: https://r4ds.had.co.nz/

WEKSLER G., KOZLOWSKI D., SHOKIDA N., "Curso de R para procesamiento de datos de la Encuesta Permanente de Hogares" (2018). Recuperado de: https://diegokoz.github.io/Curso_R_EPH_clases/

MANGINI F., "6 TIPS TO MAKE YOUR VISUALIZATIONS LOOK PROFESSIONAL [UPDATED]" (2018). Recuperado de: http://www.thinkingondata.com/6-tips-to-make-your-visualizations-look-professional/

THE R GRAPH GALLERY. "GENERAL GGPLOT2 TIPS". Recuperado de: https://www.r-graph-gallery.com/portfolio/ggplot2-package/