Lectura de datos

Maximiliano Cacace

24/5/2021

Leer los datos

```
library(readxl)
Becarios_2020 <- read_excel("C:/Users/hp/Desktop/TODO/UNSAM/Doctorado Ciencias Humanas/Métodos cuantitat
#View(Becarios_2020)

Becarios_2021 <- read_excel("C:/Users/hp/Desktop/TODO/UNSAM/Doctorado Ciencias Humanas/Métodos cuantitat
```

#“‘{r CSV} library(readr) Becarios_2020 <- read_delim(“C:/Users/hp/Desktop/TODO/UNSAM/Doctorado Ciencias Humanas/Métodos cuantitativos y análisis de grandes datos/BIG_DATA_CACACE/Becarios 2020.csv”, “;”, escape_double = FALSE, trim_ws = TRUE) View(Becarios_2020)

Becarios_2021 <- read_delim(“C:/Users/hp/Desktop/TODO/UNSAM/Doctorado Ciencias Humanas/Métodos cuantitativos y análisis de grandes datos/BIG_DATA_CACACE/Becarios 2021.csv”, “;”, escape_double = FALSE, trim_ws = TRUE) View(Becarios_2021) #“‘

```
str(Becarios_2020)
```

```
## tibble [165 x 30] (S3: tbl_df/tbl/data.frame)
##  $ Edad                                                                : num [1:165] 22 37 22 22 1
##  $ Nacionalidad                                                        : chr [1:165] "Argentina" "
##  $ DNI                                                                 : num [1:165] 40738524 2978
##  $ Carrera_que_cursa                                                   : chr [1:165] "Ingeniería c
##  $ Tiempo_de_beca                                                      : chr [1:165] "Este es el p
##  $ Costeo_de_estudios                                                  : chr [1:165] "Aportes fami
##  $ Con_quien_vive                                                      : chr [1:165] "Solo" "Famil
##  $ Por_que_eligio_ la_UNSAM                                            : chr [1:165] "Por el progr
##  $ Educacion_Primaria                                                  : chr [1:165] "Estatal" "Es
##  $ Educacion_Secundaria                                                : chr [1:165] "Estatal" "Es
##  $ Relacion_con_la_carrera                                             : chr [1:165] "SI" "NO" "SI
##  $ Repitencia_o_abandono_temporal_de_estudios                          : chr [1:165] "No" "No" "No
##  $ Tiempo_transcurrido_desde_el secundario_hasta_ingresar_a_la_universidad: chr [1:165] "Ingresé al a
##  $ Que_hizo_en_ese_tiempo                                              : chr [1:165] "Terminé el s
##  $ Trabajo                                                             : chr [1:165] "No, por deci
##  $ Que_actividad_desempeña                                             : chr [1:165] NA "Docente p
##  $ Horas_semanales_de_trabajo                                          : chr [1:165] NA "Entre 10
##  $ Aprobacion_del_CPU                                                  : chr [1:165] "Si" "Si" "Si
##  $ Año_de_inicio_del_CPU                                               : num [1:165] 2017 2016 201
##  $ Cuatrimestre_de_inicio_CPU                                          : chr [1:165] "1er Cuatrime
##  $ Alguna_dificultad_en_el_CPU_describir_en_caso_afirmativo            : chr [1:165] "No" "No" "No
##  $ Cuantos_cuatrimestres_le_llevo_aprobar_el_CPU                       : num [1:165] 1 1 1 1 NA 1
##  $ Año_de_inicio_de_carrera                                            : num [1:165] 2017 2016 201
##  $ Cuatrimestre_de_inicio_de_carrera                                   : chr [1:165] "2do Cuatrime
```

```
##  $ Cantidad_de_materias_que cursa                                    : num [1:165] 4 4 3 4 NA 0
##  $ Finales_pendientes                                                : num [1:165] 3 1 0 0 NA 1
##  $ Seguimiento_del_plan_de_estudios                                  : chr [1:165] "Parcialmente
##  $ Dificultades_observadas_en_el_seguimiento_del_plan_de_estudios    : chr [1:165] "Acumulación
##  $ Horas_semanales_destinadas_a_estudiar                             : chr [1:165] "Más de 20 ho
##  $ Porque_decidio_estudiar_en_la_Universidad                         : chr [1:165] "Vocación" "V
```

Cambiar factor y nombre de respuesta de variable

```
Becarios_2020$Tiempo_de_beca <-
  factor(as.character(Becarios_2020$Tiempo_de_beca),
         labels = c("Este es el primer año" = "0.5",
                    "1 año" = "1",
                    "2 años" = "2",
                    "3 años" = "3",
                    "4 años" = "4")
  )

Becarios_2020$Tiempo_de_beca <-
  as.numeric(Becarios_2020$Tiempo_de_beca)
```

```
Becarios_2020$Tiempo_de_beca %>%
  class()
```

```
## [1] "numeric"
```

```
table(Becarios_2020$Tiempo_de_beca)
```

```
##
##  1  2  3  4  5
## 31 17 14  6 97
```

```
Becarios_2020$Cuatrimestre_de_inicio_CPU <-
  factor(as.character(Becarios_2020$Cuatrimestre_de_inicio_CPU),
         labels = c("1er Cuatrimestre" = "1",
                    "2do Cuatrimestre" = "2")
  )

Becarios_2020$Cuatrimestre_de_inicio_CPU <-
  factor(as.numeric(Becarios_2020$Cuatrimestre_de_inicio_CPU),
         labels = c("1" = "1",
                    "2" = "2")
  )
Becarios_2020$Cuatrimestre_de_inicio_CPU %>%
  class()
```

```
## [1] "factor"
```

```
table(Becarios_2020$Cuatrimestre_de_inicio_CPU)
```

```
##
##   1   2
## 132  33
```

```
Becarios_2020$Cuatrimestre_de_inicio_CPU %>%
  class()
```

```
## [1] "factor"
```

```
table(Becarios_2020$Cuatrimestre_de_inicio_CPU)
```

```
##
##   1   2
## 132  33
```

```
nrow(Becarios_2020)
```

```
## [1] 165
```

```
ncol(Becarios_2020)
```

```
## [1] 30
```

```
nrow(Becarios_2021)
```

```
## [1] 151
```

```
ncol(Becarios_2021)
```

```
## [1] 30
```

```
#Ratio = 8
#Nominal = 22

Becarios_2020$Edad %>% #Ratio
  class()
```

```
## [1] "numeric"
```

```
Becarios_2020$Nacionalidad %>% #Nominal
  class()
```

```
## [1] "character"
```

```
Becarios_2020$DNI %>% #Nominal
  class()
```

```
## [1] "numeric"
```

```
Becarios_2020$Carrera_que_cursa %>% #Nominal
  class()
```

```
## [1] "character"
```

```
Becarios_2020$Tiempo_de_beca %>% #Ratio
  class()
```

```
## [1] "numeric"
```

```
Becarios_2020$Costeo_de_estudios %>% #Nominal
  class()
```

```
## [1] "character"
```

```
Becarios_2020$Con_quien_vive %>% #Nominal
  class()
```

```
## [1] "character"
```

```
Becarios_2020$`Por_que_eligio_ la_UNSAM` %>% #Nominal
  class()
```

```
## [1] "character"
```

```
Becarios_2020$Educacion_Primaria %>% #Nominal
  class()
```

```
## [1] "character"
```

```
Becarios_2020$Educacion_Secundaria %>% #Nominal
  class()
```

```
## [1] "character"
```

```
Becarios_2020$Relacion_con_la_carrera %>% #Nominal
  class()
```

```
## [1] "character"
```

```
Becarios_2020$Repitencia_o_abandono_temporal_de_estudios %>% #Nominal
  class()
```

```
## [1] "character"
```

```
Becarios_2020$`Tiempo_transcurrido_desde_el secundario_hasta_ingresar_a_la_universidad` %>% #Ratio
  class()
```

```
## [1] "character"
```

```
Becarios_2020$Que_hizo_en_ese_tiempo %>% #Nominal
  class()
```

```
## [1] "character"
```

```
Becarios_2020$Trabajo %>% #Nominal
  class()
```

```
## [1] "character"
```

```
Becarios_2020$Que_actividad_desempeña %>% #Nominal
  class()
```

```
## [1] "character"
```

```
Becarios_2020$Horas_semanales_de_trabajo %>% #Ratio
  class()
```

```
## [1] "character"
```

```
Becarios_2020$Aprobacion_del_CPU %>% #Nominal
  class()
```

```
## [1] "character"
```

```
Becarios_2020$Año_de_inicio_del_CPU %>% #Nominal
  class()
```

```
## [1] "numeric"
```

```
Becarios_2020$Cuatrimestre_de_inicio_CPU %>% #Nominal
  class()
```

```
## [1] "factor"
```

```
Becarios_2020$Alguna_dificultad_en_el_CPU_describir_en_caso_afirmativo %>% #Nominal
  class()
```

```
## [1] "character"
```

```
Becarios_2020$Cuantos_cuatrimestres_le_llevo_aprobar_el_CPU %>% #Ratio
  class()
```

```
## [1] "numeric"
```

```
Becarios_2020$Año_de_inicio_de_carrera %>% #Nominal
  class()
```

```
## [1] "numeric"
```

```
Becarios_2020$Cuatrimestre_de_inicio_de_carrera %>% #Nominal
  class()
```

```
## [1] "character"
```

```
Becarios_2020$`Cantidad_de_materias_que cursa` %>% #Ratio
  class()
```

```
## [1] "numeric"
```

```
Becarios_2020$Finales_pendientes %>% #Ratio
  class()
```

```
## [1] "numeric"
```

```
Becarios_2020$Seguimiento_del_plan_de_estudios %>% #Nominal
  class()
```

```
## [1] "character"
```

```
Becarios_2020$Dificultades_observadas_en_el_seguimiento_del_plan_de_estudios %>% #Nominal
  class()
```

```
## [1] "character"
```

```
Becarios_2020$Horas_semanales_destinadas_a_estudiar %>% #Ratio
  class()
```

```
## [1] "character"
```

```
Becarios_2020$Porque_decidio_estudiar_en_la_Universidad %>% #Nominal
  class()
```

```
## [1] "character"
```

```
Becarios_2020$Edad
```

```
##   [1] 22 37 22 22 26 25 25 21 26 22 20 26 20 36 18 20 17 19 26 34 21 22 19 25 25
##  [26] 18 22 59 32 25 22 20 26 21 29 22 20 27 33 23 22 19 22 45 28 32 33 25 60 30
##  [51] 27 23 24 24 18 18 24 26 48 27 25 19 20 25 25 29 22 25 18 26 29 32 22 28 21
##  [76] 23 22 22 19 22 20 20 23 24 42 24 18 25 29 25 19 33 22 43 27 23 26 35 24 20
## [101] 32 20 22 21 20 20 21 25 23 22 22 19 30 20 17 25 37 19 20 20 21 23 37 18 28
## [126] 31 23 20 19 19 20 32 23 20 26 22 25 27 26 21 27 21 25 23 23 35 24 20 20 22
## [151] 20 33 34 36 19 19 25 23 26 20 25 18 19 19 24
```

```
Becarios_2021$Edad
```

```
##   [1] 26 23 22 22 26 55 21 27 18 21 37 21 20 44 27 24 24 23 29 26 36 23 21 48 28
##  [26] 23 28 23 17 40 19 42 46 29 33 34 26 19 30 36 31 20 24 25 25 20 49 28 18 30
##  [51] 23 20 26 30 26 23 19 21 27 30 32 23 25 38 22 23 20 26 26 21 20 21 25 27 43
##  [76] 26 20 22 26 34 45 24 27 33 36 26 25 21 21 23 21 22 26 42 24 19 23 35 31 22
```

```
## [101] 23 25 38 20 23 28 21 22 25 38 19 25 22 34 23 19 24 51 21 26 41 23 33 22 27
## [126] 39 23 18 28 22 20 22 26 23 36 21 25 25 21 23 20 23 19 29 34 20 21 23 22 24
## [151] 29
```

```
table(Becarios_2020$Edad)
```

```
##
## 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 42 43 45 48 59
##  2  8 14 22  9 21 12  8 18 11  6  3  4  2  1  5  4  2  2  2  3  1  1  1  1  1
## 60
##  1
```

```
table(Becarios_2021$Edad)
```

```
##
## 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42
##  1  3  7 11 15 12 20  7 10 14  6  5  4  4  2  1  3  4  1  4  1  3  1  1  1  2
## 43 44 45 46 48 49 51 55
##  1  1  1  1  1  1  1  1
```

```
table(Becarios_2020$Carrera_que_cursa)
```

```
##
##                     Ingeniería Ambiental                     Ingeniería Biomédica
##                                       20                                       12
##                   Ingeniería Electrónica                    Ingeniería en Energía
##                                       10                                       10
##         Ingeniería en Telecomunicaciones                      Ingeniería Espacial
##                                        4                                        3
##                    Ingeniería Industrial            Licenciatura en Biotecnología
##                                        8                                       40
##            Licenciatura en Física Médica Tecnicatura en Diagnóstico por Imágenes
##                                        4                                       38
##           Tecnicatura en Electromedicina Tecnicatura en Programación Informática
##                                        2                                       12
##        Tecnicatura en Redes Informáticas
##                                        2
```

```
Becarios_2020 %>%
   group_by(Costeo_de_estudios) %>%
   summarize(
     mean(Edad),
     median(Edad)
   )
```

```
## # A tibble: 5 x 3
##   Costeo_de_estudios                          `mean(Edad)` `median(Edad)`
##   <chr>                                              <dbl>          <dbl>
## 1 Aportes familiares                                  22.4           22
## 2 Beca                                                23.8           22
## 3 Con dificultad para costear sus estudios            26.9           25.5
## 4 Plan social                                         60             60
## 5 Trabajo                                             27.4           24.5
```

```
mean(Becarios_2020$Edad) -> promedio_Edad2020
promedio_Edad2020
```

```
## [1] 24.86061
```

```r
mean(Becarios_2021$Edad) -> promedio_Edad2021
promedio_Edad2021
```

```
## [1] 26.66887
```

```r
median(Becarios_2020$Edad) -> median_Edad2020
median_Edad2020
```

```
## [1] 23
```

```r
median(Becarios_2021$Edad) -> median_Edad2021
median_Edad2021
```

```
## [1] 24
```

```r
range(Becarios_2020$Edad) #Rango o Amplitud de la variable (valor máx. y min.)
```

```
## [1] 17 60
```

```r
quantile(Becarios_2020$Edad) #Cuantiles, deciles y percentiles (se dividen las obs. en 4 partes iguales
```

```
##   0%  25%  50%  75% 100%
##   17   20   23   26   60
```

```r
IQR(Becarios_2020$Edad) #Rango intercuartil (diferencia entre el tercer y el primer cuartil de una dist
```

```
## [1] 6
```

```r
var(Becarios_2020$Edad) #Varianza
```

```
## [1] 46.53533
```

```r
sd(Becarios_2020$Edad) #Desviación Estándar
```

```
## [1] 6.821681
```

```r
mad(Becarios_2020$Edad) #Desviación Mediana Absoluta
```

```
## [1] 4.4478
```

```r
summary(Becarios_2020$Edad)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   20.00   23.00   24.86   26.00   60.00
```

```r
pnorm(Becarios_2020$Edad)
```

```
##   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [75] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [112] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [149] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

```r
order(Becarios_2020$Edad, decreasing=FALSE)
```

```
##   [1]  17 115  15  26  55  56  69  87 124 162  18  23  42  62  79  91 112 118
##  [19] 129 130 155 156 163 164  11  13  16  32  37  63  81  82 100 102 105 106
##  [37] 114 119 120 128 131 134 148 149 151 160   8  21  34  75 104 107 121 140
##  [55] 142   1   3   4  10  22  27  31  36  41  43  67  73  77  78  80  93 103
##  [73] 110 111 136 150  40  52  76  83  96 109 122 127 133 144 145 158  53  54
##  [91]  57  84  86  99 147 165   6   7  24  25  30  48  61  64  65  68  88  90
## [109] 108 116 137 143 157 161   5   9  12  19  33  58  70  97 135 139 159  38
```

```
## [127]  51  60  95 138 141  45  74 125  35  66  71  89  50 113 126  29  46  72
## [145] 101 132  39  47  92 152  20 153  98 146  14 154   2 117 123  85  94  44
## [163]  59  28  49
```

Tests

```
shapiro.test(Becarios_2020$Edad)
```

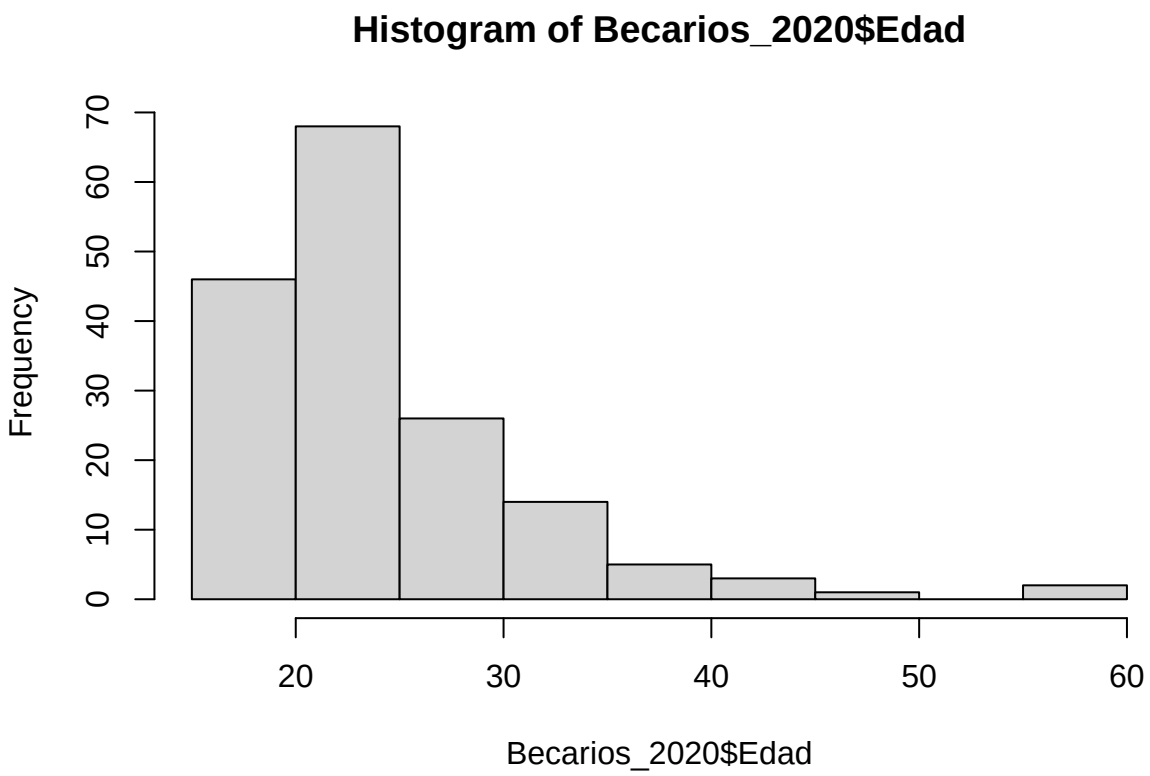
```
##
##  Shapiro-Wilk normality test
##
## data:  Becarios_2020$Edad
## W = 0.78464, p-value = 2.625e-14
```

```
t.test(Becarios_2020$Edad, Becarios_2021$Edad)
```

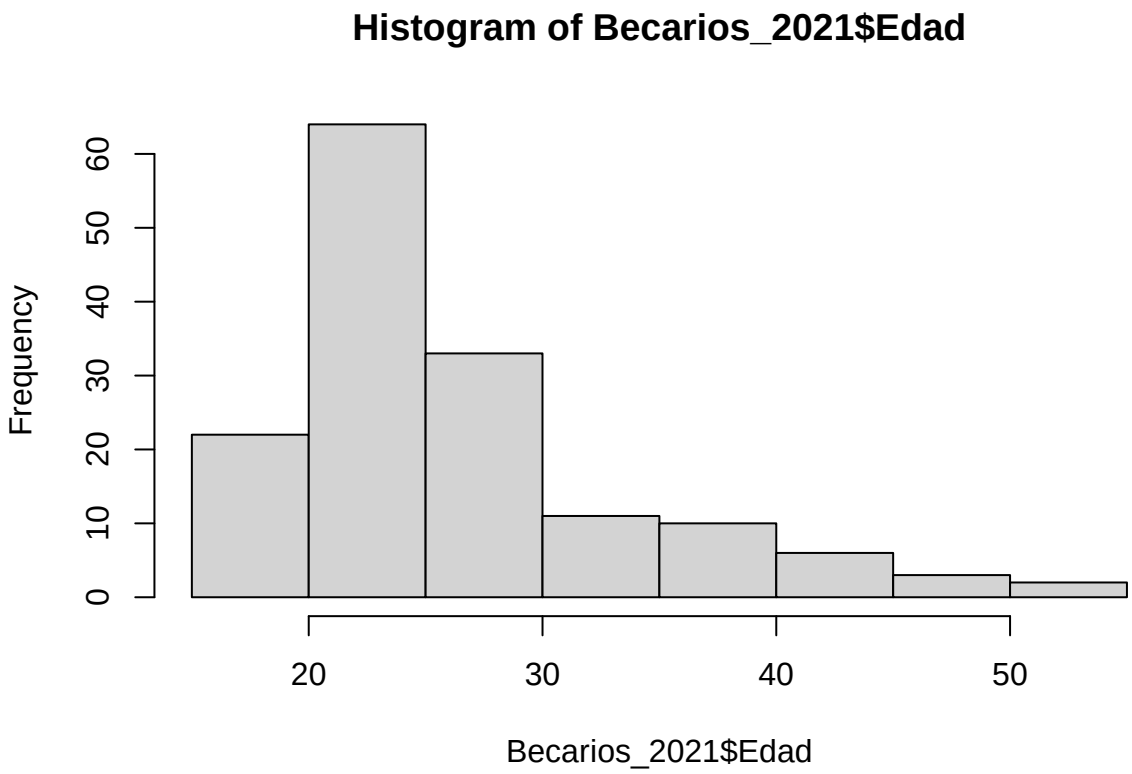
```
##
##  Welch Two Sample t-test
##
## data:  Becarios_2020$Edad and Becarios_2021$Edad
## t = -2.243, df = 304.47, p-value = 0.02562
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.3946927 -0.2218436
## sample estimates:
## mean of x mean of y
##  24.86061  26.66887
```

##Visualización de datos

```
hist(Becarios_2020$Edad)
```

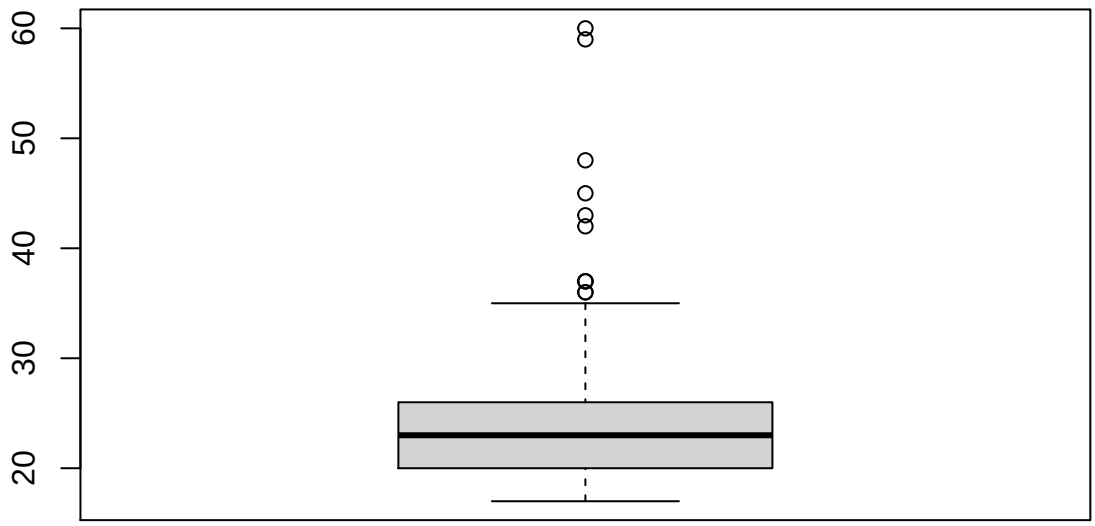



```
hist(Becarios_2021$Edad)
```

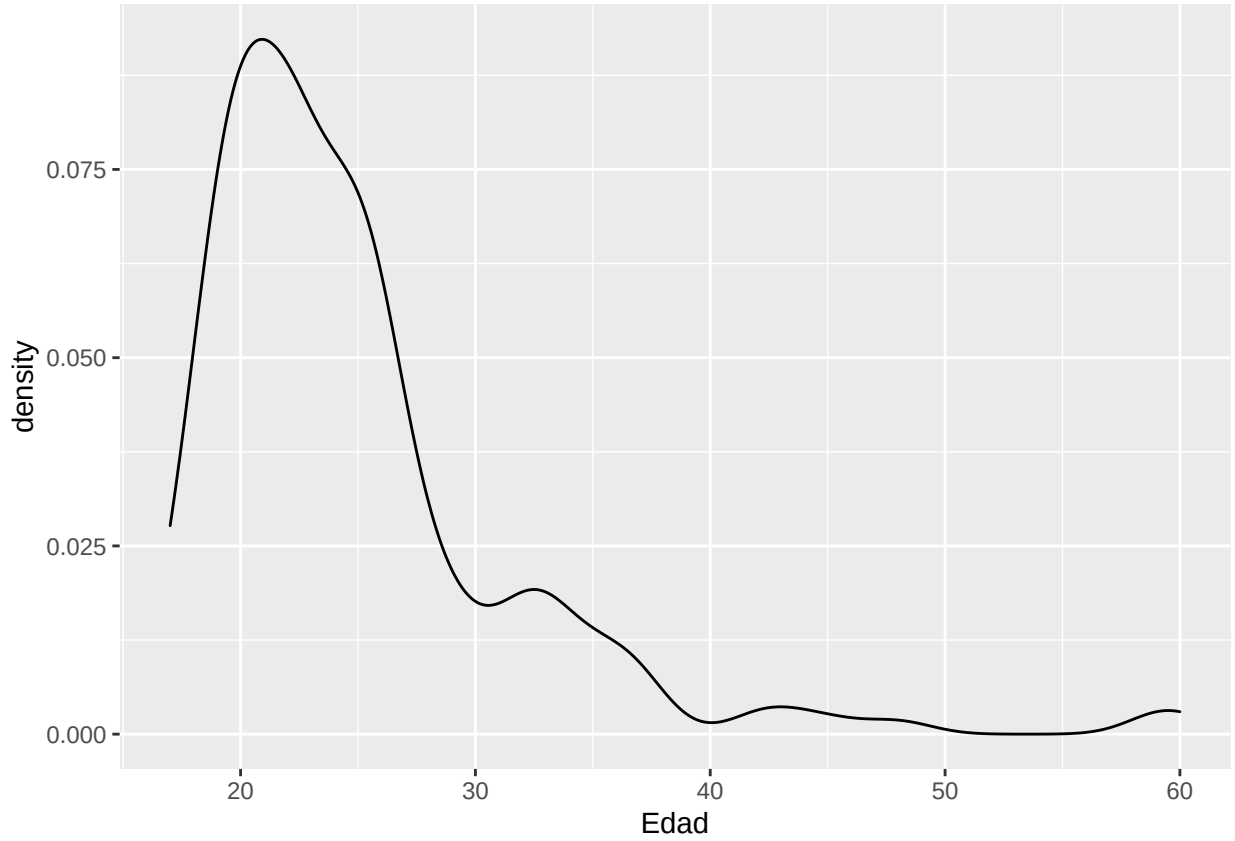



```
#polygon(Becarios_2020$Edad)
```

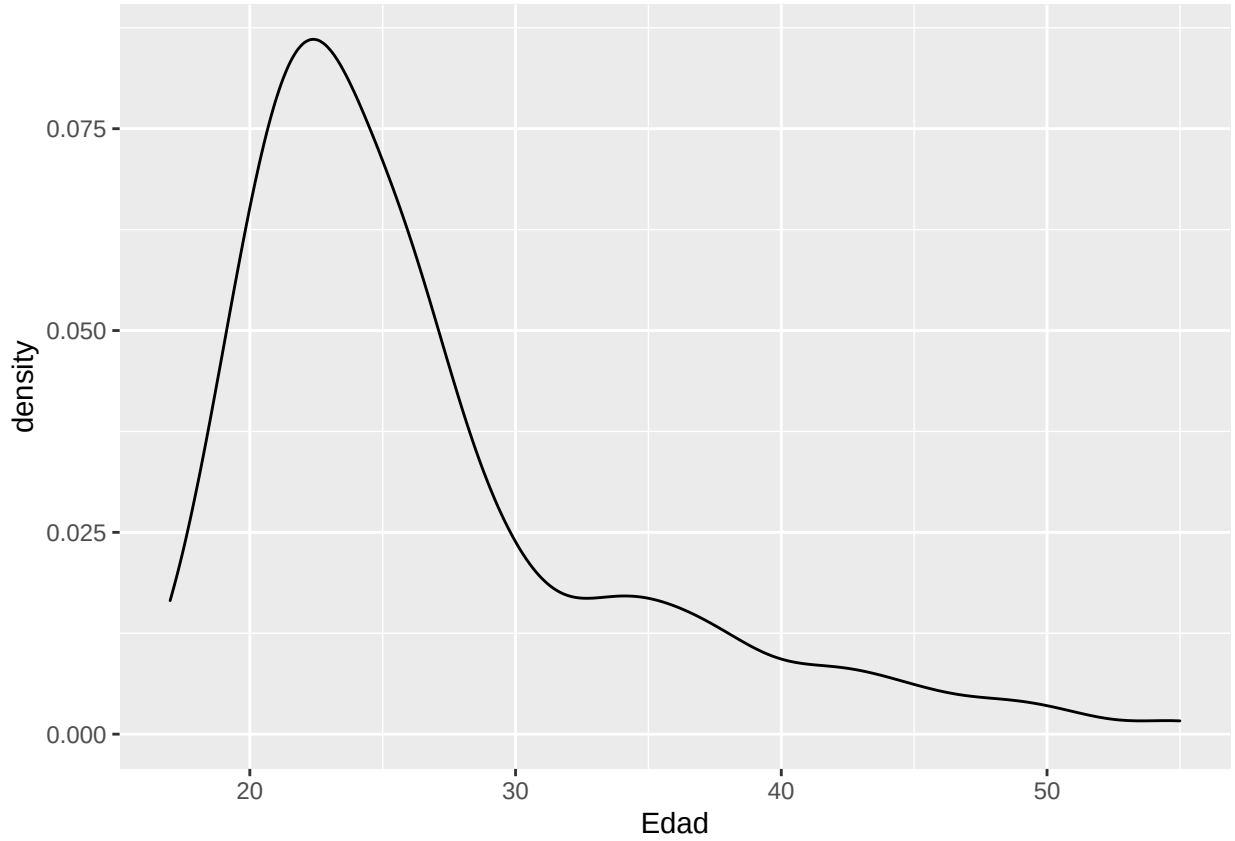
```
boxplot(Becarios_2020$Edad) #visualización de la dispersión
```



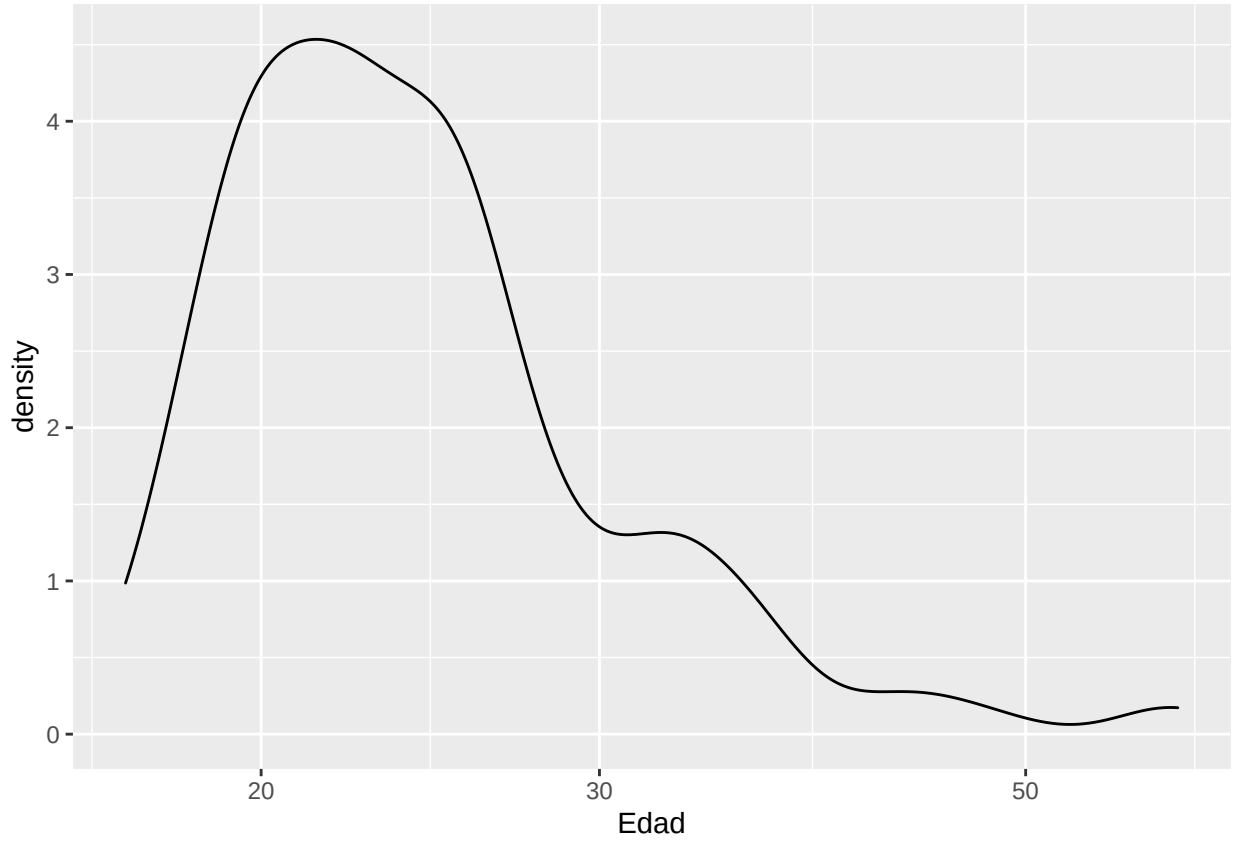

```
library(ggplot2)
Becarios_2020 %>%
  ggplot(aes(Edad))+
  geom_density()
```




```
Becarios_2021 %>%
  ggplot(aes(Edad))+
  geom_density()
```

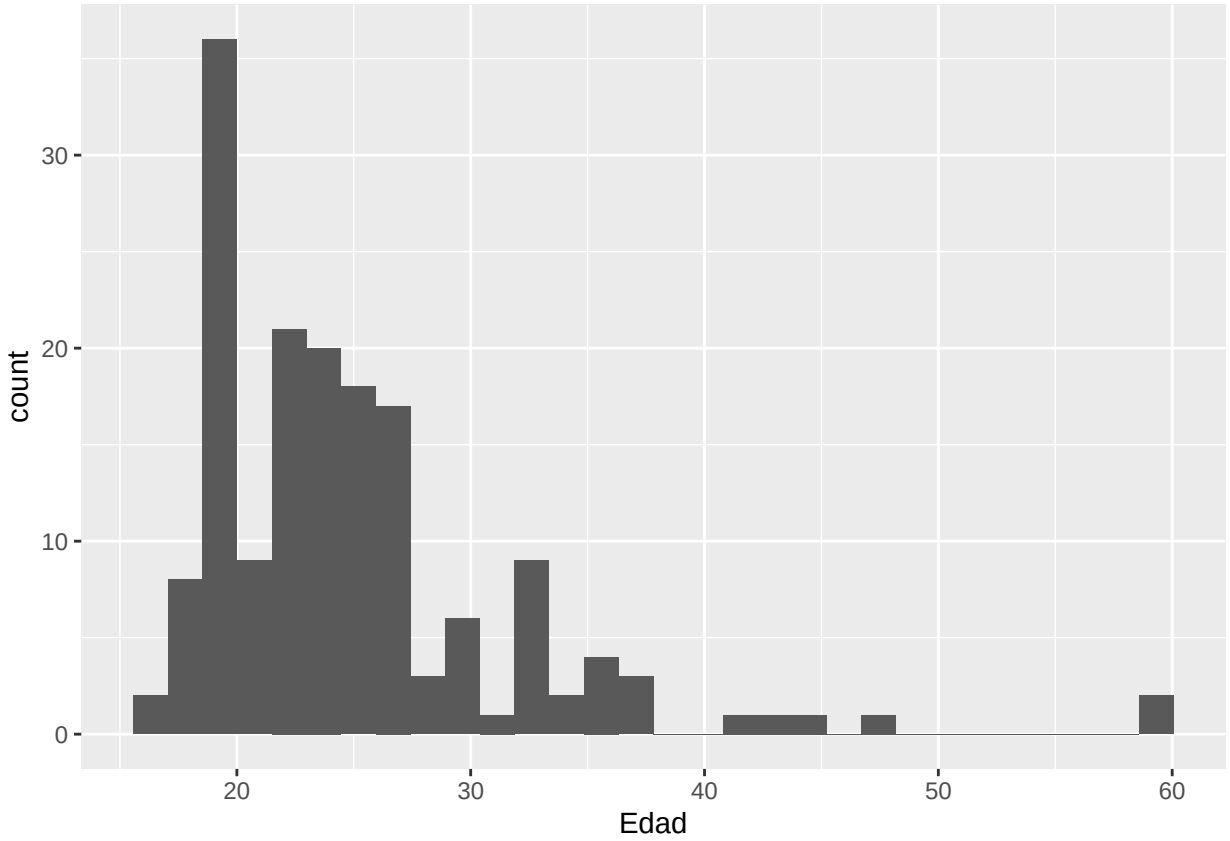



```
library(ggplot2)
Becarios_2020 %>%
  ggplot(aes(Edad))+
  geom_density()+
  scale_x_log10()
```



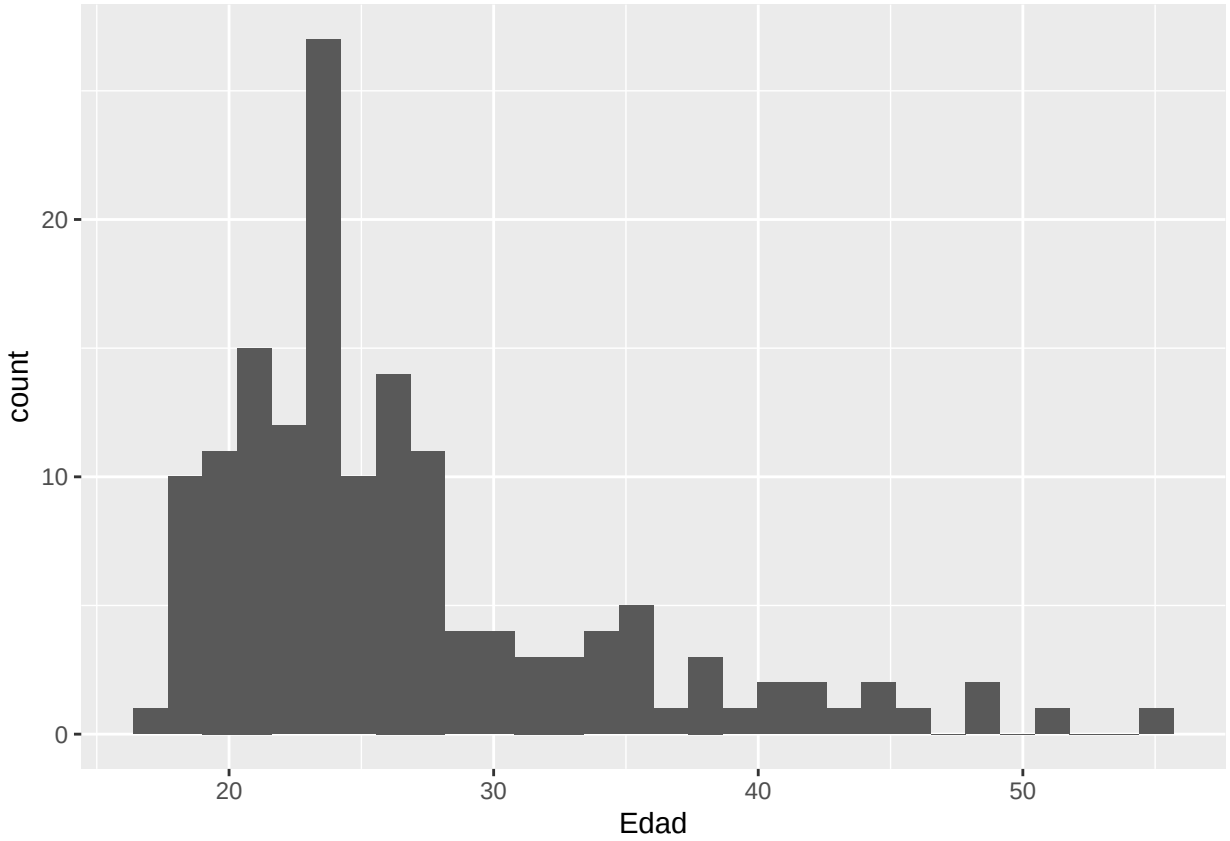

```
library(ggplot2)
options(scipen=100)
Becarios_2020 %>%
  ggplot(aes(x=Edad))+
  geom_histogram()
```

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

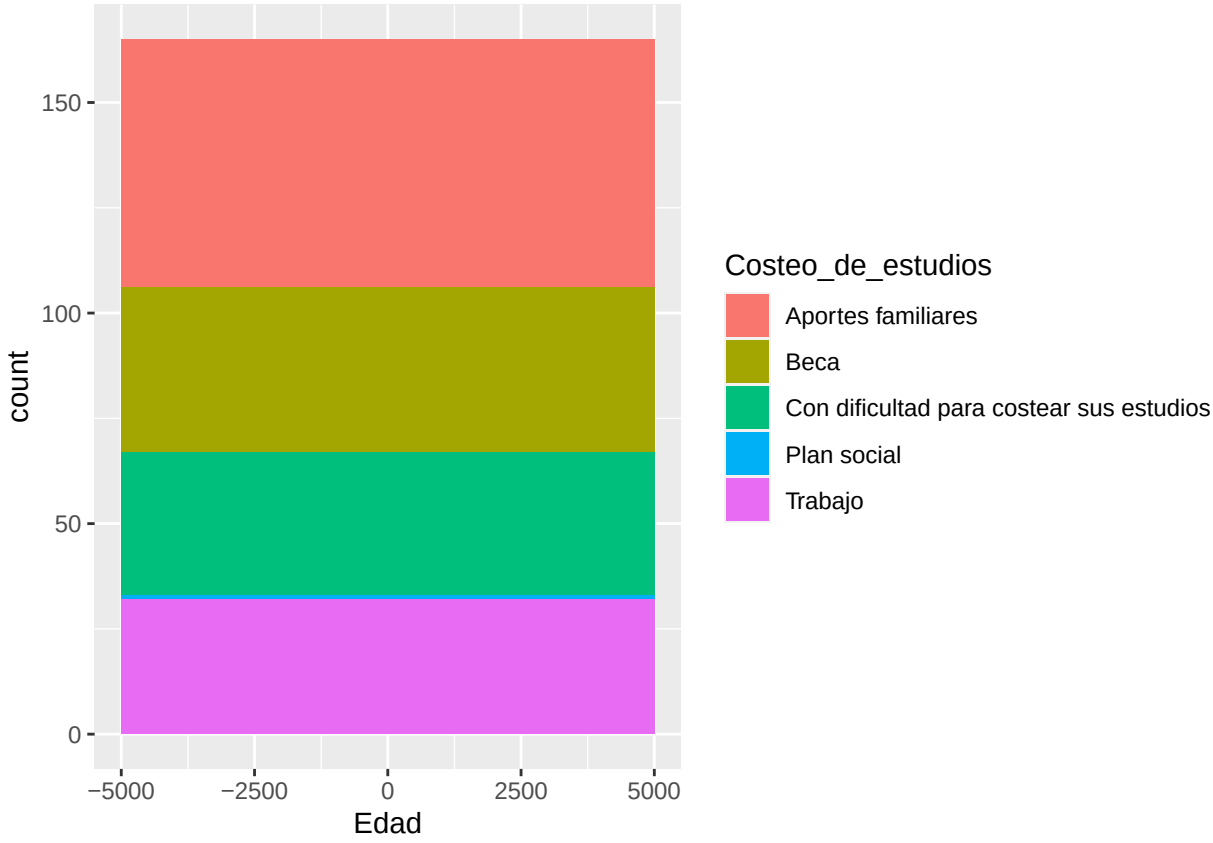



```
Becarios_2021 %>%
  ggplot(aes(x=Edad))+
  geom_histogram()
```

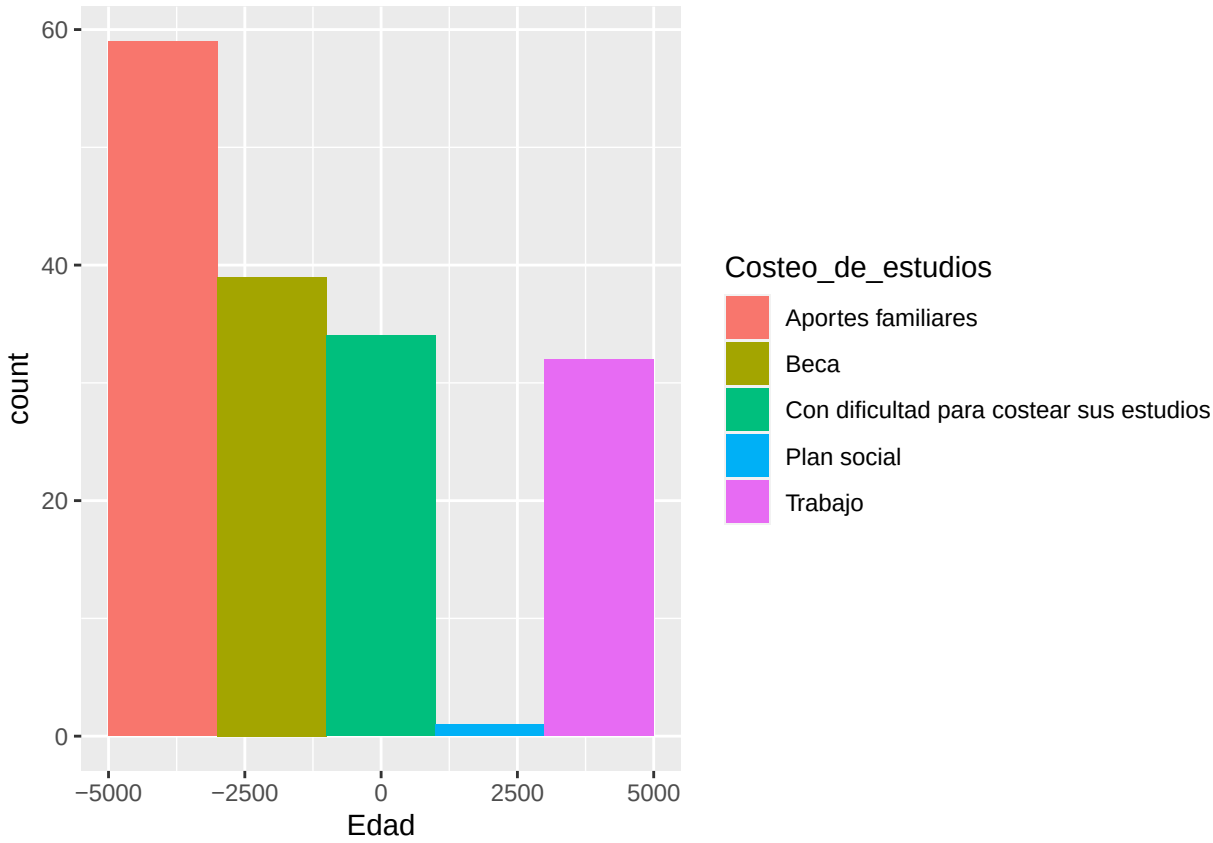
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



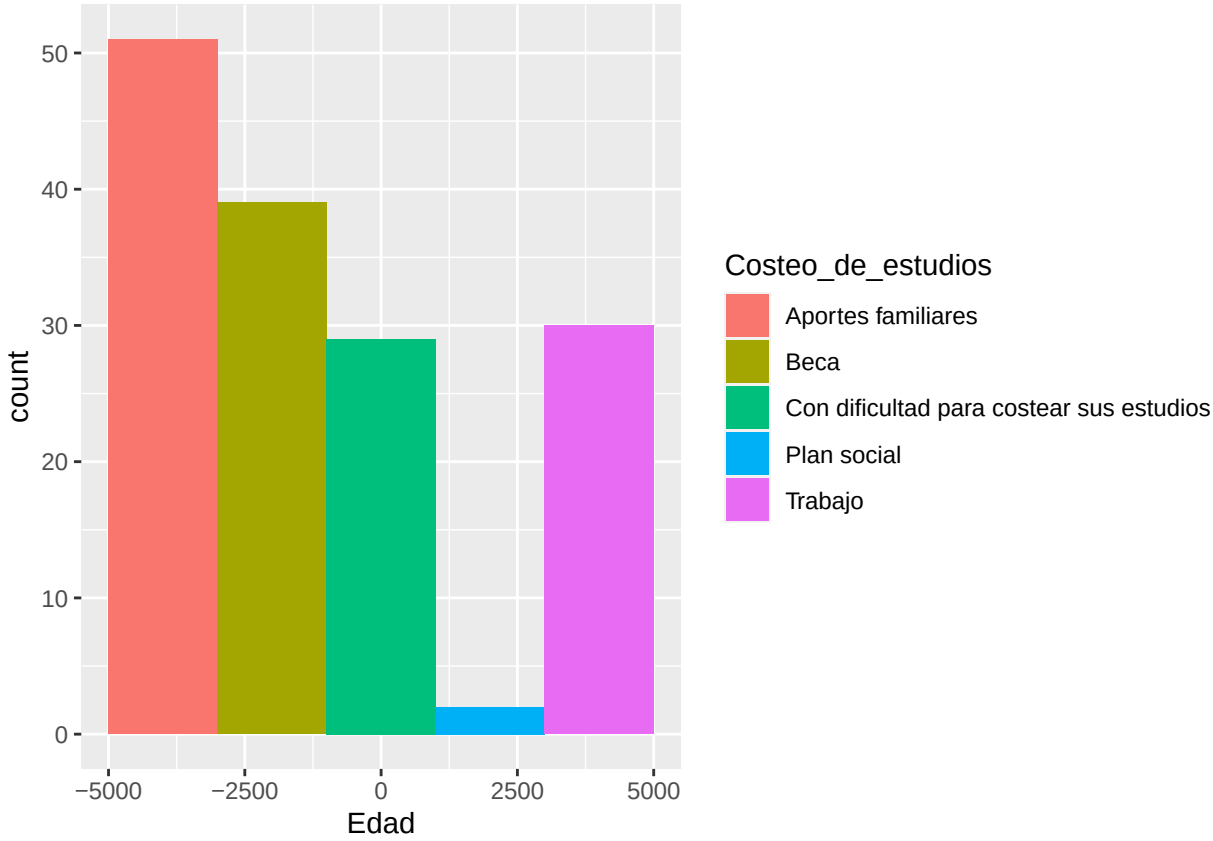

```
library(ggplot2)
Becarios_2020 %>%
  ggplot(aes(x=Edad, fill=Costeo_de_estudios))+
  geom_histogram(binwidth = 10000)
```




```
Becarios_2020 %>%
  ggplot(aes(x=Edad, fill=Costeo_de_estudios))+
  geom_histogram(binwidth = 10000, position = "dodge")
```



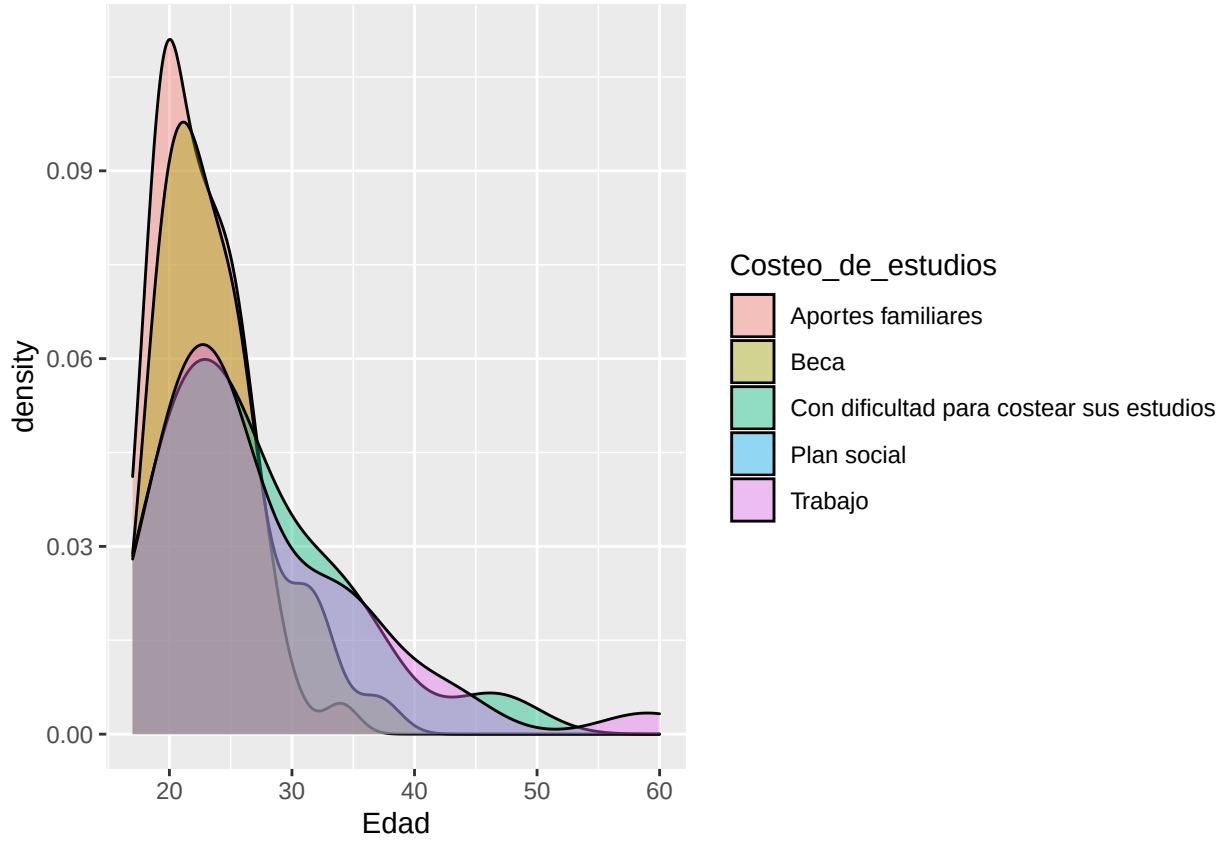

```
Becarios_2021 %>%
  ggplot(aes(x=Edad, fill=Costeo_de_estudios))+
  geom_histogram(binwidth = 10000, position = "dodge")
```



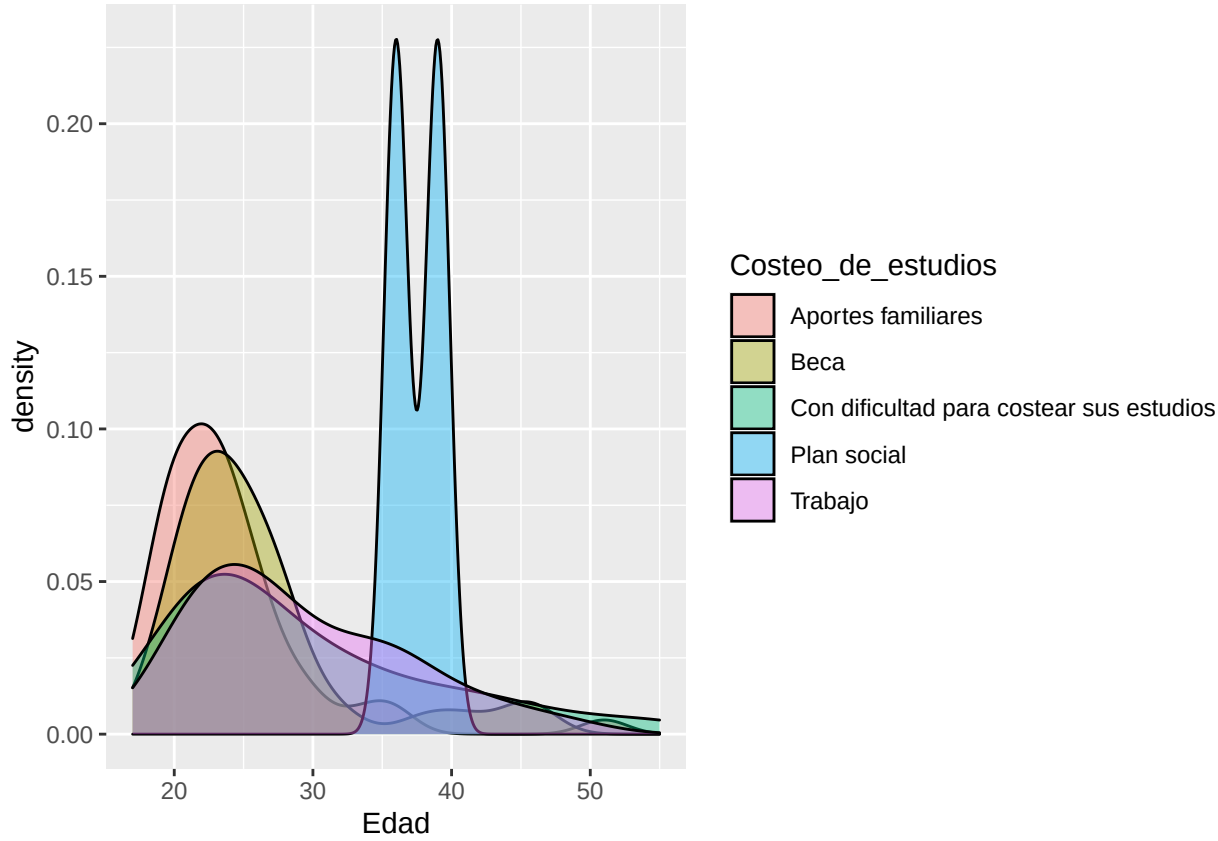

```
Becarios_2020%>%
  ggplot(aes(x=Edad, fill=Costeo_de_estudios))+
  geom_density(alpha=.4)
```

```
## Warning: Groups with fewer than two data points have been dropped.
```

```
## Warning in max(ids, na.rm = TRUE): ningun argumento finito para max; retornando
## -Inf
```




```
Becarios_2021%>%
  ggplot(aes(x=Edad, fill=Costeo_de_estudios))+
  geom_density(alpha=.4)
```




```
#library(ggplot2)
#p <- ggplot(Becarios_2020)
#p <- p + aes (x = Edad, y = Costeo_de_estudios)
#p <- p + geom_point()
#p
```