

TP #3

ESTADÍSTICA - UTN FRSR '22

ANA ASCUA

TRABAJO PRACTICO NRO 3

LENGUAJE R

PRIMER ETAPA

- 1- Generar en Excel una base de datos que tenga al menos 50 registros o mas, como minimo 5 campos, con datos numéricos y alfanuméricos.
- 2- Generar luego en R-ESTUDIO, un script que muestre dicho archivo (en la solapa de script) captura de pantalla.
- 3- En la solapa de variables y observaciones mostrar las variables del script (captura de pantalla)
- 4- Instalar en la solapa inferior derecha las librerías tidy, readxl, ggplot2, dplyr
- 5- Con dos de las columnas de la tabla seleccionada generar un grafico usando ggplot2 y mostrar graficos de puntos, barra, lineal y fraccionado. (captura de pantalla)
- 6- Instalar una nueva librería que se llama tidyverse (captura de pantalla)
- 7- Usando esta librería calcular la media, mediana y desviación estándar. (en la solapa de consola) captura de pantalla del uso de la librería y mostrar resultados.

TRABAJO PRACTICO NRO 3 SEGUNDA ETAPA

NOMBRE DE EQUIPO:

GGPLOT

- 1- CON LOS EJEMPLOS DE DATOS DEL PRACTICO ANTERIOR Y UTILIZANDO GGPLOT
 - A- GRAFICAR EN : BARPLOT, BARPLOT EN COLORES, BARPLOT CON ORIENTACIONES CAMBIADAS, CON COLORES DISTINTOS PARA CADA VARIABLE.
 - B- HISTOGRAMA: HISTOGRAMA EN LA CUAL CADA COLUMNA TENGA UN RANGO DE 0 A 2, HACER UN HISTOGRAMA HACIENDO CORTES, HACER UN HISTOGRAMA USANDO LA VARIABLE CARAT, HACER UN HISTOGRAMA USANDO LA VARIABLE CUT, HACER UN HISTOGRAMA USANDO LA VARIABLE CUT Y COLOR.
 - C- GRAFICOS DE DISPERSION: GRAFICOS DE PUNTOS

ETAPA 1:

1) BASE DE DATOS: Genero una base de datos con **5 campos**, ID, nombre, edad, teléfono y mail.

Algunos de los datos son generados aleatoriamente y de manera automatizada en Sheets de Google

docs.google.com/spreadsheets/d/1hf18KL7SZkQraoSx2MiiCo-A5vVUgHB2vKZzzcddCY/edit#gid=0

base de datos 50 registros

Archivo Editar Ver Insertar Formato Datos Herramientas Extensiones Ayuda Última modificación hace 6 días

75% € % .0 .00 123 Arial 12 B I S A

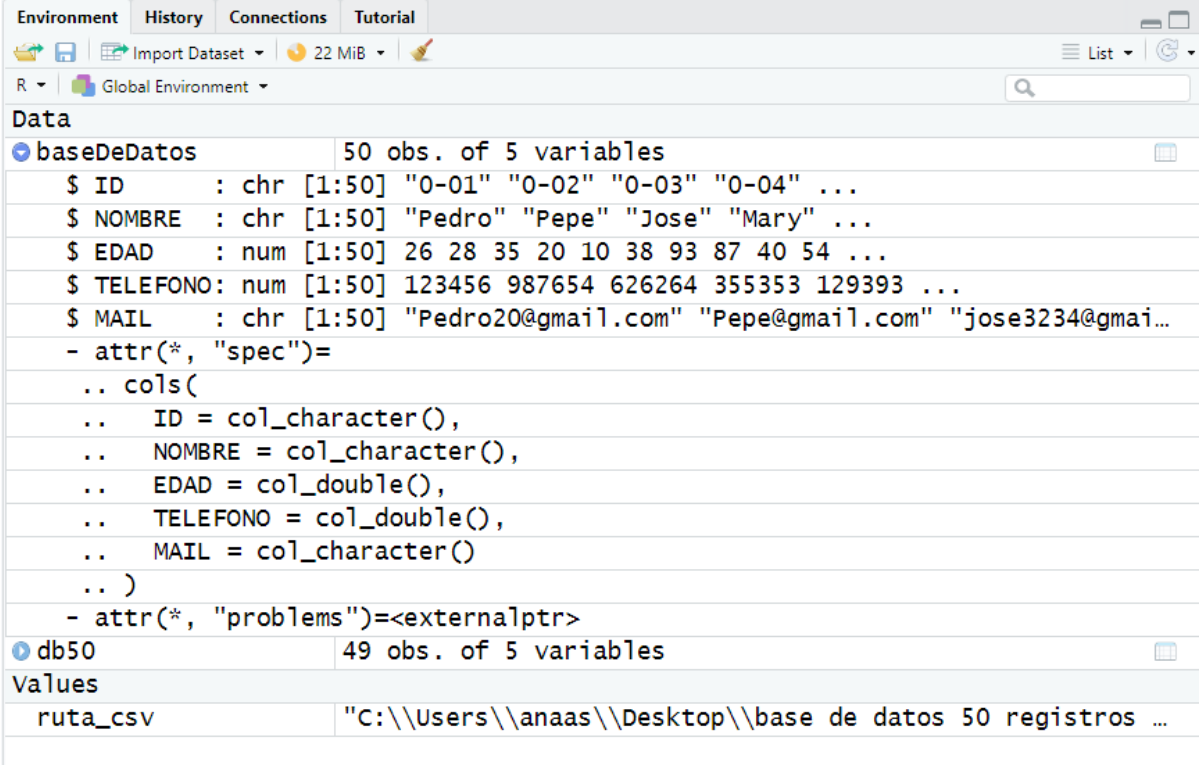
A1	ID																
	ID	NOMBRE	EDAD	TELEFONO	MAIL												
1	0-01	Pedro	26	123456	Pedro20@gmail.com												
2	0-02	Pepe	28	987654	Pepe@gmail.com												
3	0-03	Jose	35	626264	jose3234@gmail.com												
4	0-04	Mary	20	355353	mary@gmail.com												
5	0-05	Gio	10	129393	gio@gmail.com												
6	0-06	Antonia	17	789023	Antonia@gmail.com												
7	0-07	Wanda	51	3694566	Wanda@gmail.com												
8	0-08	Dua Lipa	38	4204207	Dua Lipa@gmail.com												
9	0-09	Mario	65	136790	Mario@gmail.com												
10	0-10	Jack	45	135679	Jack@gmail.com												
11	0-11	Juan	62	876592	Juan@gmail.com												
12	0-12	Luciano	66	9488324	Luciano@gmail.com												
13	0-13	Leandro	43	8795624	Leandro@gmail.com												
14	0-14	Fernando	98	4497886	Fernando@gmail.com												
15	0-15	Marcos	26	9795624	Marcos@gmail.com												
16	0-16	Facundo	15	9798653	Facundo@gmail.com												
17	0-17	Daniel	41	6498675	Daniel@gmail.com												
18	0-18	Hector	82	2854945	Hector@gmail.com												
19	0-19	Nacho	28	5485644	Nacho@gmail.com												
20	0-20	Francisco	37	6485646	Francisco@gmail.com												
21	0-21	Roberto	47	5184684	Roberto@gmail.com												
22	0-22	Rafael	48	5482846	Rafael@gmail.com												
23	0-23	Juan Cruz	89	8485643	Juan Cruz@gmail.com												
24	0-24	Lina	43	8487906	Lina@gmail.com												
25	0-25	Jose	43	9765353	Jose@gmail.com												
26	0-26	Luisa	53	7676835	Luisa@gmail.com												
27	0-27	Maria Jose	47	3234984	Maria Jose@gmail.com												
28	0-28	Roberta	56	1446785	Roberta@gmail.com												
29	0-29	Lelia	79	4666863	Lelia@gmail.com												
30	0-30	Paula	69	4815162	Paula@gmail.com												
31	0-31	Jonathan	40	1623427	Jonathan@gmail.com												
32	0-32	Marisa	63	4873268	Marisa@gmail.com												
33	0-33	Mariana	93	8465314	Mariana@gmail.com												
34	0-34	Karina	88	1346706	Karina@gmail.com												

+ Hoja 1

2) Script en RStudio:

```
1 # cargamos el paquete tidyverse (incluye ggplot2)
2 # library(tidyverse)
3
4 # buscamos la ruta del archivo de csv
5 file.choose()
6
7 # copiamos la ruta de la consola y la guardamos en una variable
8 ruta_csv <- "C:\\Users\\anaas\\Desktop\\base de datos 50 registros - Hoja 1 (1).csv"
9
10 # importamos los datos
11 # yo le puse de nombre baseDeDatos con camelCase (la base de datos tiene 50 registros)
12 baseDeDatos <- read_csv(ruta_csv)
13
14 # miramos los datos con la función View (V mayúscula)
15 View(baseDeDatos)
16
17 # graficamos con GGplot2
18 ggplot(data = baseDeDatos, aes(x = EDAD)) + geom_bar()
19
20 # -----|
21 # X es un eje
22 # Y es otro eje
23
20:28 (Untitled) R Script
```

3) Variables del script:



The screenshot shows the RStudio Environment pane with the following content:

Environment | History | Connections | Tutorial

Import Dataset | 22 MiB

R | Global Environment

Data

baseDeDatos 50 obs. of 5 variables

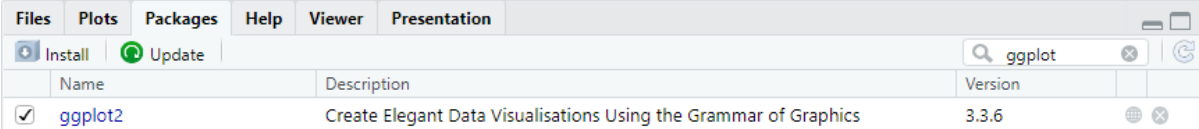
```
$ ID      : chr [1:50] "0-01" "0-02" "0-03" "0-04" ...
$ NOMBRE  : chr [1:50] "Pedro" "Pepe" "Jose" "Mary" ...
$ EDAD    : num [1:50] 26 28 35 20 10 38 93 87 40 54 ...
$ TELEFONO: num [1:50] 123456 987654 626264 355353 129393 ...
$ MAIL    : chr [1:50] "Pedro20@gmail.com" "Pepe@gmail.com" "jose3234@gmai..."
- attr(*, "spec")=
.. cols(
..   ID = col_character(),
..   NOMBRE = col_character(),
..   EDAD = col_double(),
..   TELEFONO = col_double(),
..   MAIL = col_character()
.. )
- attr(*, "problems")=<externalptr>
```

db50 49 obs. of 5 variables

Values

ruta_csv "C:\\Users\\anaas\\Desktop\\base de datos 50 registros ..."

4) Instalación de librerías tidy, readxl, ggplot2, dplyr



The screenshot shows the RStudio Packages pane with the following content:

Files | Plots | Packages | Help | Viewer | Presentation

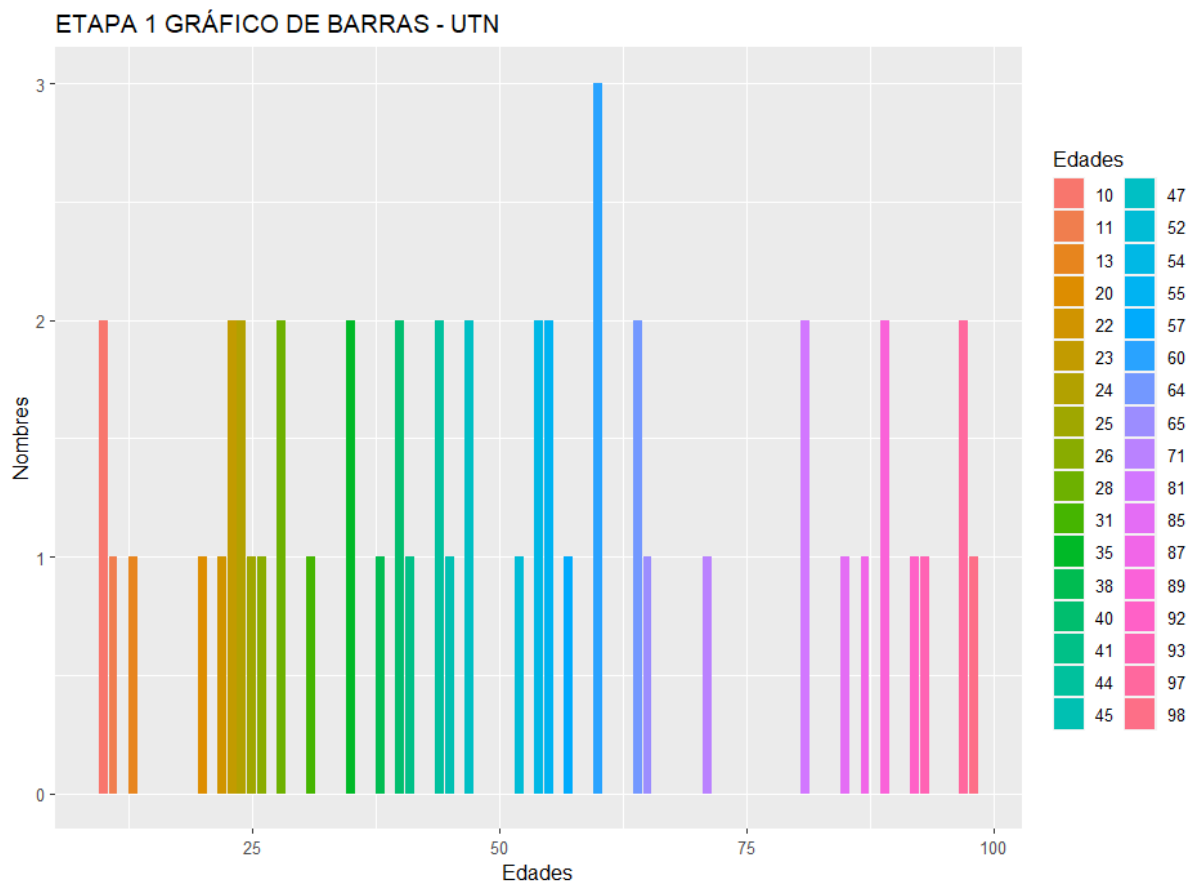
Install | Update

Search: ggplot

	Name	Description	Version
<input checked="" type="checkbox"/>	ggplot2	Create Elegant Data Visualisations Using the Grammar of Graphics	3.3.6

5) Con 2 de las columnas de la tabla seleccionada, generar un gráfico usando ggplot2 y mostrar: Gráfico de barras, Gráfico de puntos, Gráfico lineal y Gráfico fraccionado (captura de todos)

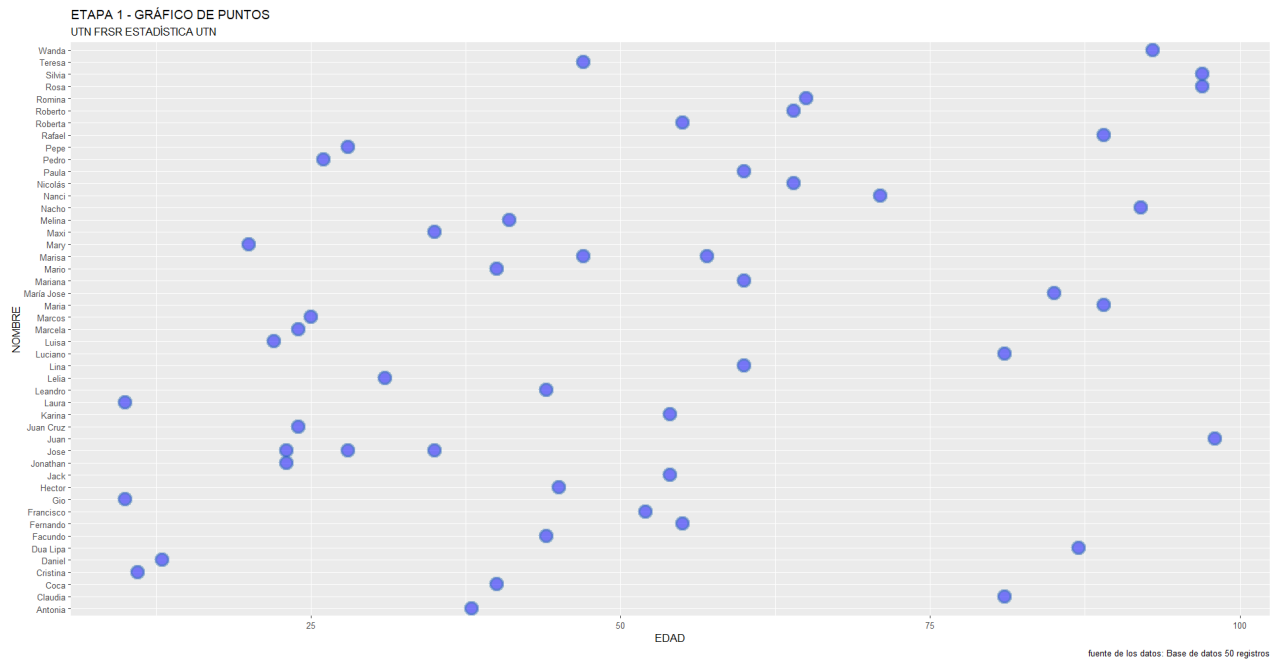
GRÁFICO DE BARRAS:



SCRIPT DEL GRÁFICO DE BARRAS Y SUS CAPAS:

```
62
63 # le asigno colores distintos a un grupo de datos (edades en este caso)
64 ggplot(data = baseDeDatos, aes(x = EDAD, fill = as.factor(EDAD))) +
65   geom_bar() +
66   xlab("Edades") +
67   ylab("Nombres") +
68   ggtitle("ETAPA 1 GRÁFICO DE BARRAS - UTN") +
69   labs(fill = "Edades")
```

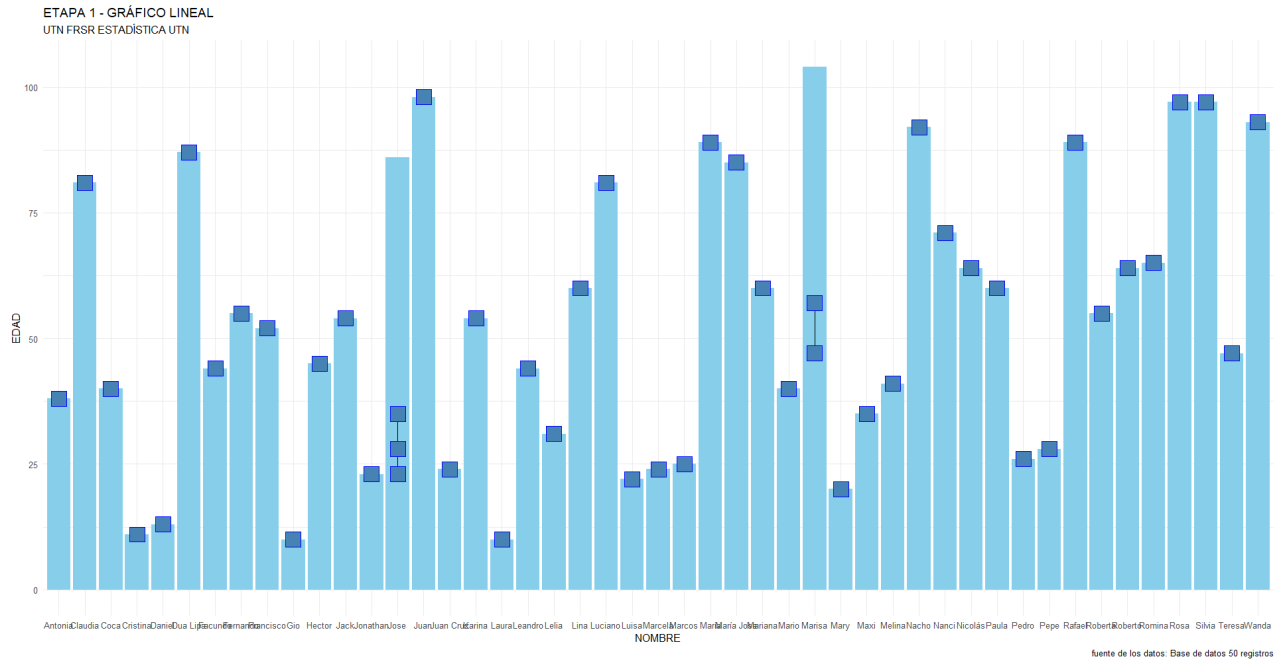
GRÁFICO DE PUNTOS:



SCRIPT DEL GRÁFICO DE PUNTOS:

```
134 # GEOM_POINT
135 ggplot(data = baseDeDatos2, mapping =
136       aes(x = EDAD, y = NOMBRE))+
137   geom_point(color = 'steelblue', fill = "blue",
138             shape = 21,
139             alpha = 0.5,
140             size = 5,
141             stroke = 2) +
142   labs(title = 'ETAPA 1 - GRÁFICO DE PUNTOS',
143        subtitle = 'UTN FRSR ESTADÍSTICA UTN',
144        caption = 'fuente de los datos: Base de datos 50 registros')
```

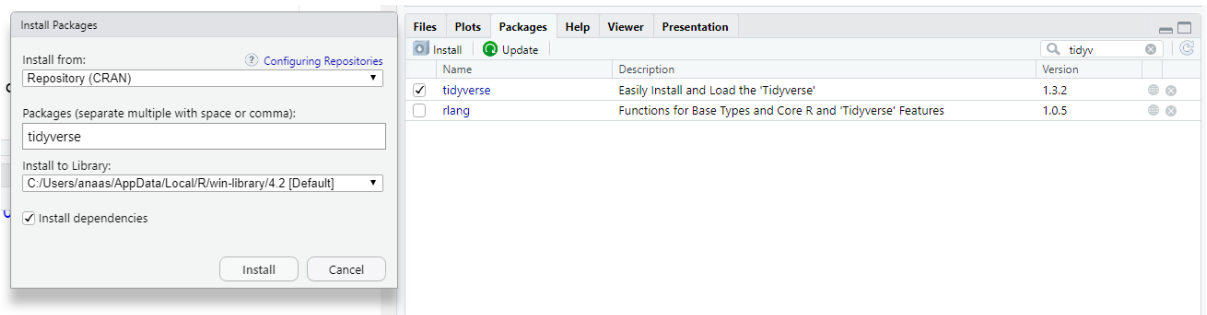
GRÁFICO LINEAL:



SCRIPT DEL GRÁFICO LINEAL:

```
148 # GRÁFICO LINEAL
149 library(ggplot2)
150
151 ggplot(data = baseDeDatos2, aes(x=NOMBRE,y=EDAD))+
152   geom_bar(stat = "identity", fill="skyblue")+
153   geom_line(color="black", stroke = 2) +
154   geom_point(size=8, shape=22, fill="steelblue",
155             color="blue")+
156   labs(title = "ETAPA 1 - GRÁFICO LINEAL",
157        subtitle = "UTN FRSR ESTADÍSTICA UTN",
158        caption = "fuente de los datos: Base de datos 50 registros")+
159   theme_minimal()
```

6) captura de la instalación de la librería TIDYVERSE EN PACKAGES RSTUDIO:



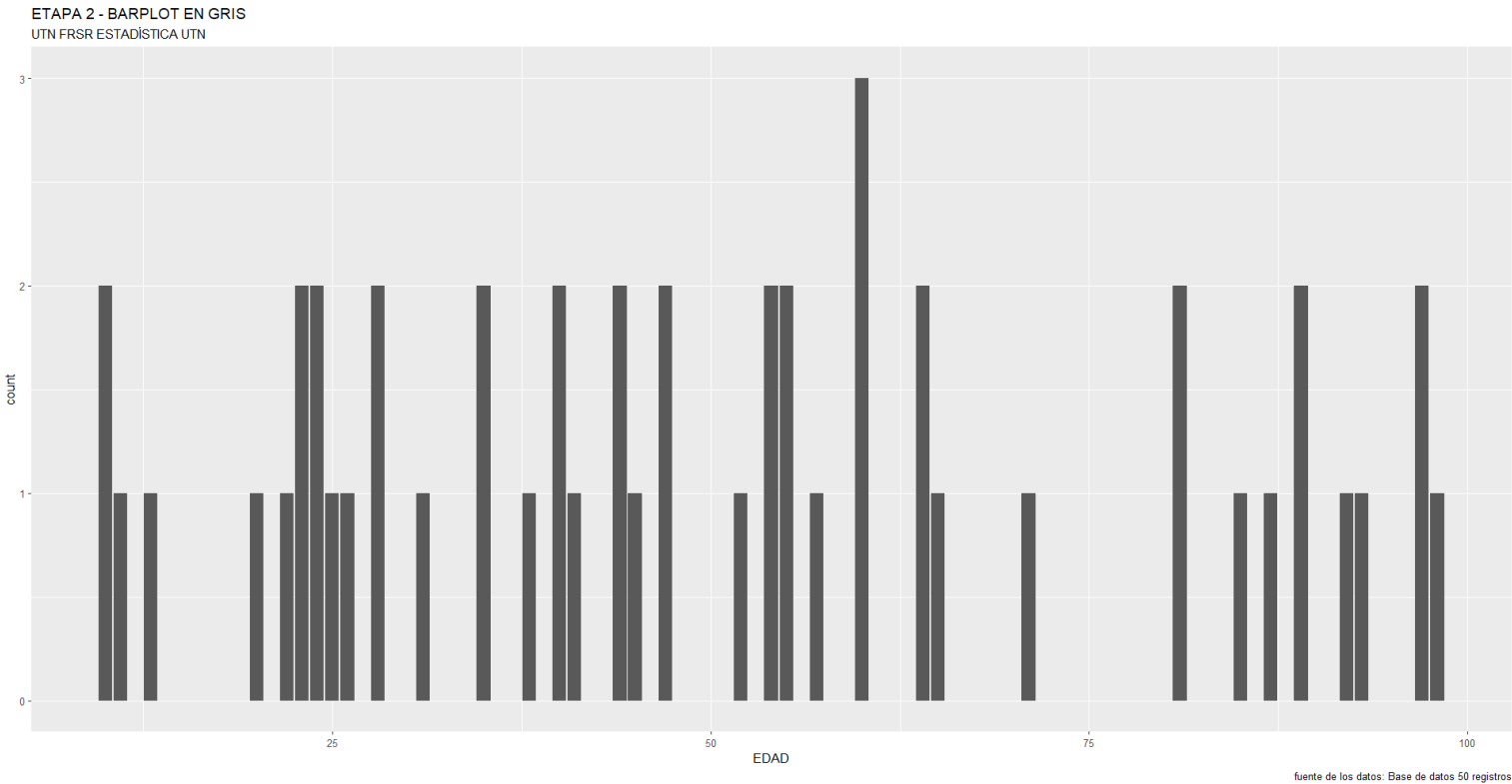
7) Calculamos la media, la mediana, la varianza y la desviación estándar con la variable \$EDAD de nuestra base de datos:

```
Console | Terminal x | Background Jobs x
R 4.2.1 · ~/

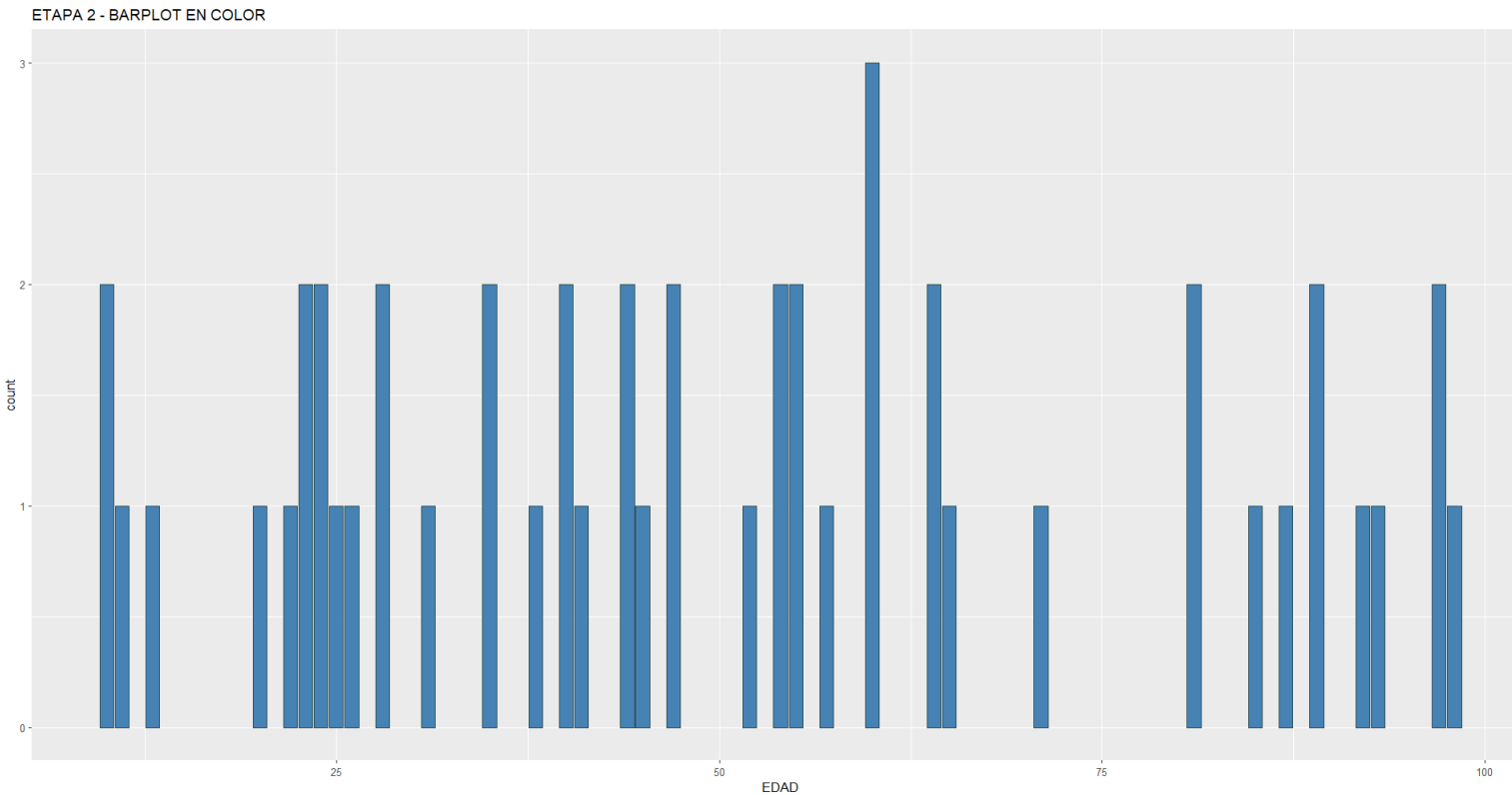
> # ETAPA 1 PUNTO 7 CON TIDIVERSE LIBRARY
> library("tidyverse")
>
> # calculamos la media de la variable EDAD de nuestra base de datos con la función mean
> mean(baseDeDatos2$EDAD)
[1] 50.68
>
> # luego calculamos la mediana con la función median
> median(baseDeDatos2$EDAD)
[1] 47
>
> # también calculamos la varianza con la función var
> var(baseDeDatos2$EDAD)
[1] 682.9976
>
> # y finalmente calculamos la desviación estándar con la función sd
> sd(baseDeDatos2$EDAD)
[1] 26.13422
> |
```

ETAPA 2:

BARPLOT EN GRIS:



BARPLOT EN COLOR:

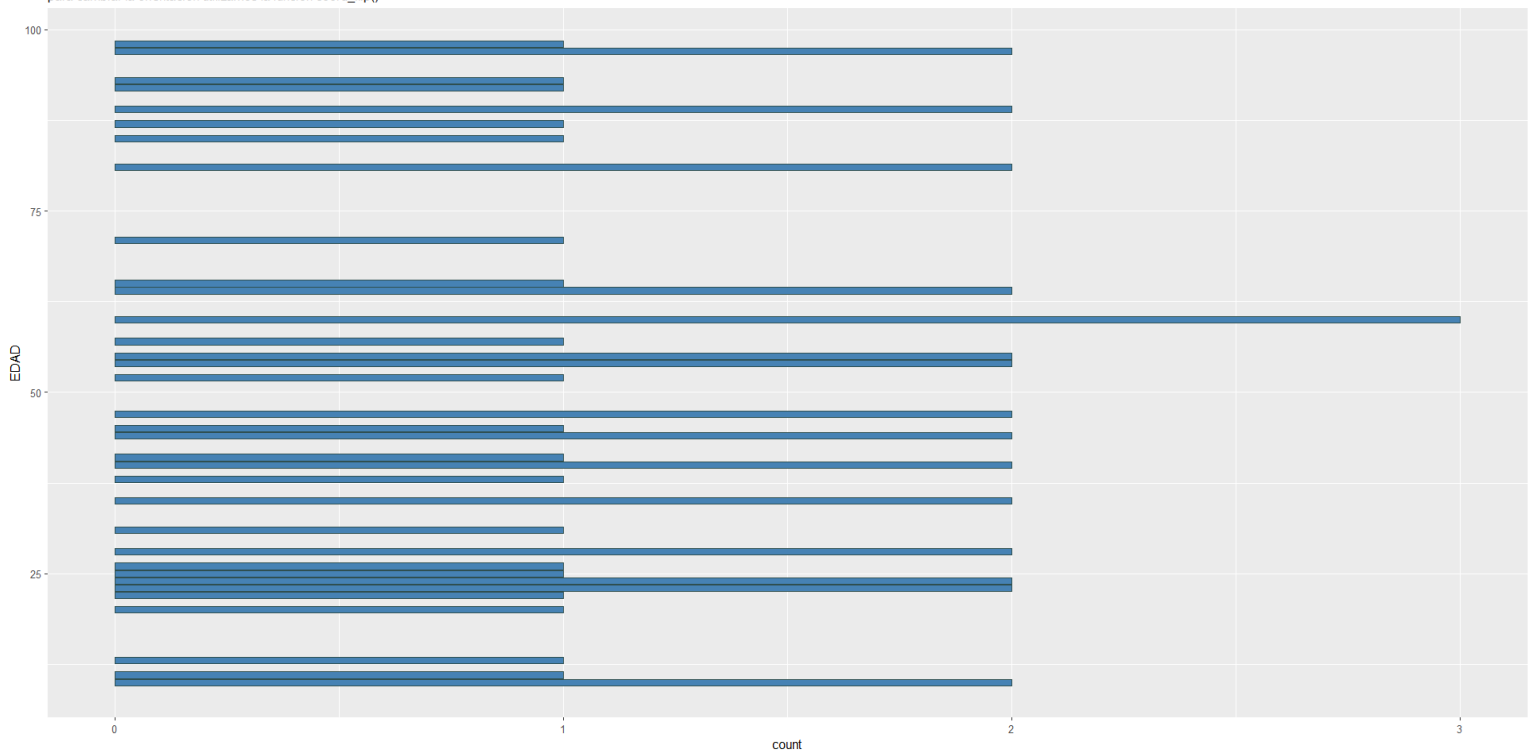


Scripts de barplot en GRIS y barplot en COLOR:

```
215 # barplot en escala de grises
216 ggplot(data = baseDeDatos2, aes(x=EDAD))+
217   geom_bar()+
218   labs(title = "ETAPA 2 - BARPLOT EN GRIS",
219         subtitle = "UTN FRSR ESTADÍSTICA UTN",
220         caption = "fuente de los datos: Base de datos 50 registros")
221
222
223 # barplot con color
224 ggplot(data = baseDeDatos2, aes(x=EDAD))+
225   geom_bar(color = 'darkslategray', fill = 'steelblue')+
226   ggtitle("ETAPA 2 - BARPLOT EN COLOR")
227
```

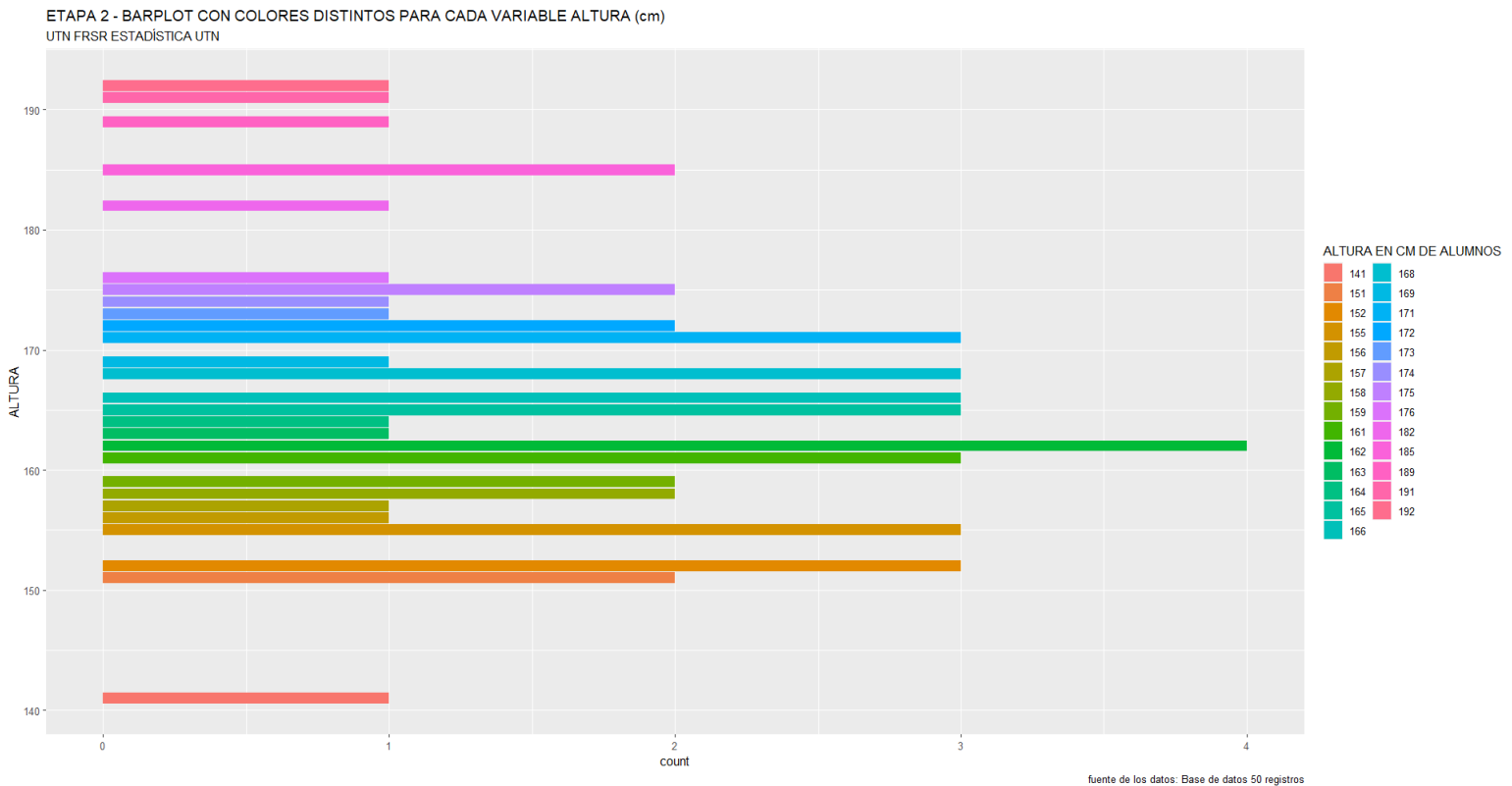
BARPLOT CON ORIENTACIONES CAMBIADAS: usando coord_flip()

ETAPA 2 - BARPLOT CON ORIENTACIONES CAMBIADAS
para cambiar la orientación utilizamos la función coord_flip()



BARPLOT CON DISTINTOS COLORES PARA CADA VARIABLE:

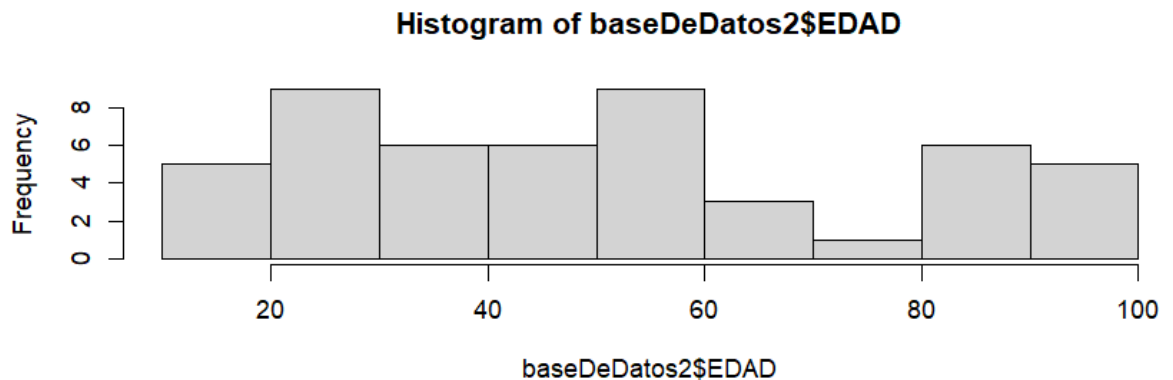
Para hacer distintos gráficos, a la base de datos le agregué 2 columnas, PESO y ALTURA.



Script del barplot con colores distintos para variable ALTURA en CM

```
246 # barplot con colores distintos para cada variable
247 ggplot(data = baseDeDatos2, aes(x = ALTURA, fill = as.factor(ALTURA))) +
248   geom_bar() +
249   xlab("ALTURA") +
250   ylab("count") +
251   labs(fill = "ALTURA EN CM DE ALUMNOS",
252        title = "ETAPA 2 - BARPLOT CON COLORES DISTINTOS PARA CADA VARIABLE ALTURA (cm)",
253        subtitle = "UTN FRSR ESTADÍSTICA UTN",
254        caption = "fuente de los datos: Base de datos 50 registros")+
255   coord_flip()
256
```

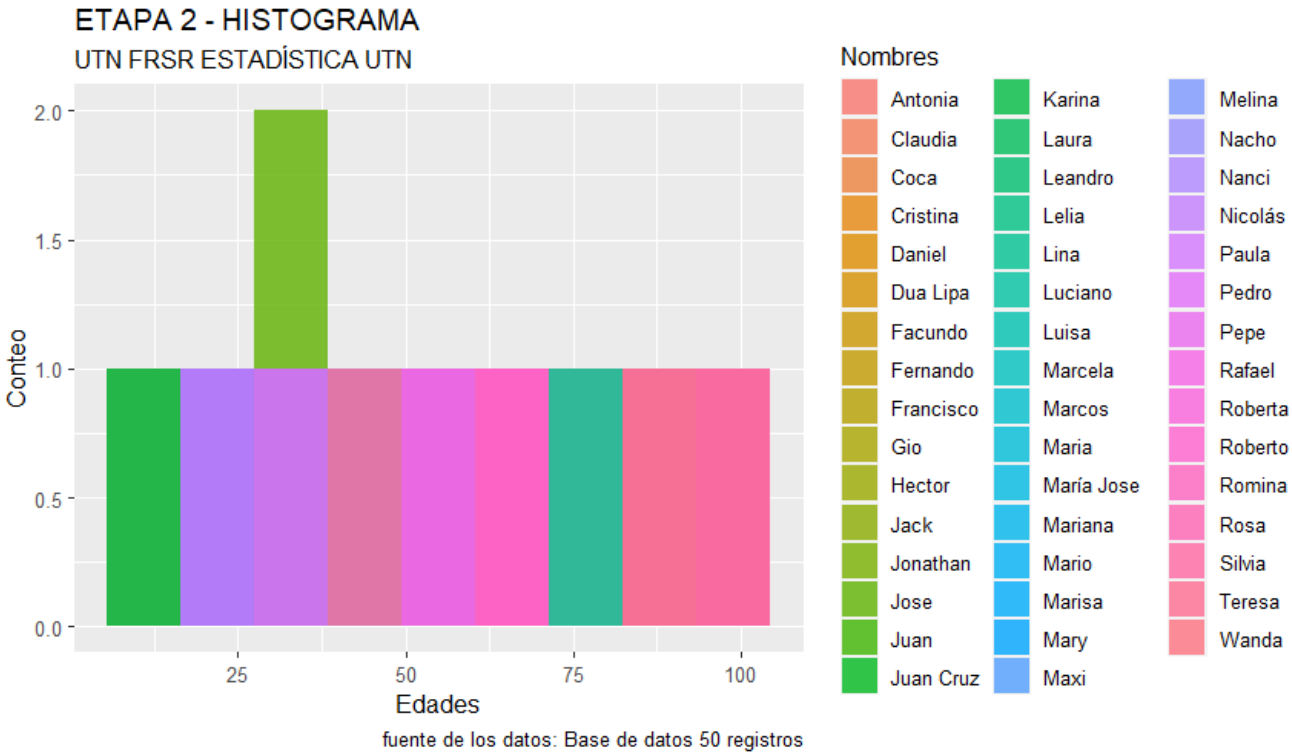
Histograma base de R, hecho con la variable \$EDAD. Nos muestra las frecuencias de las edades:



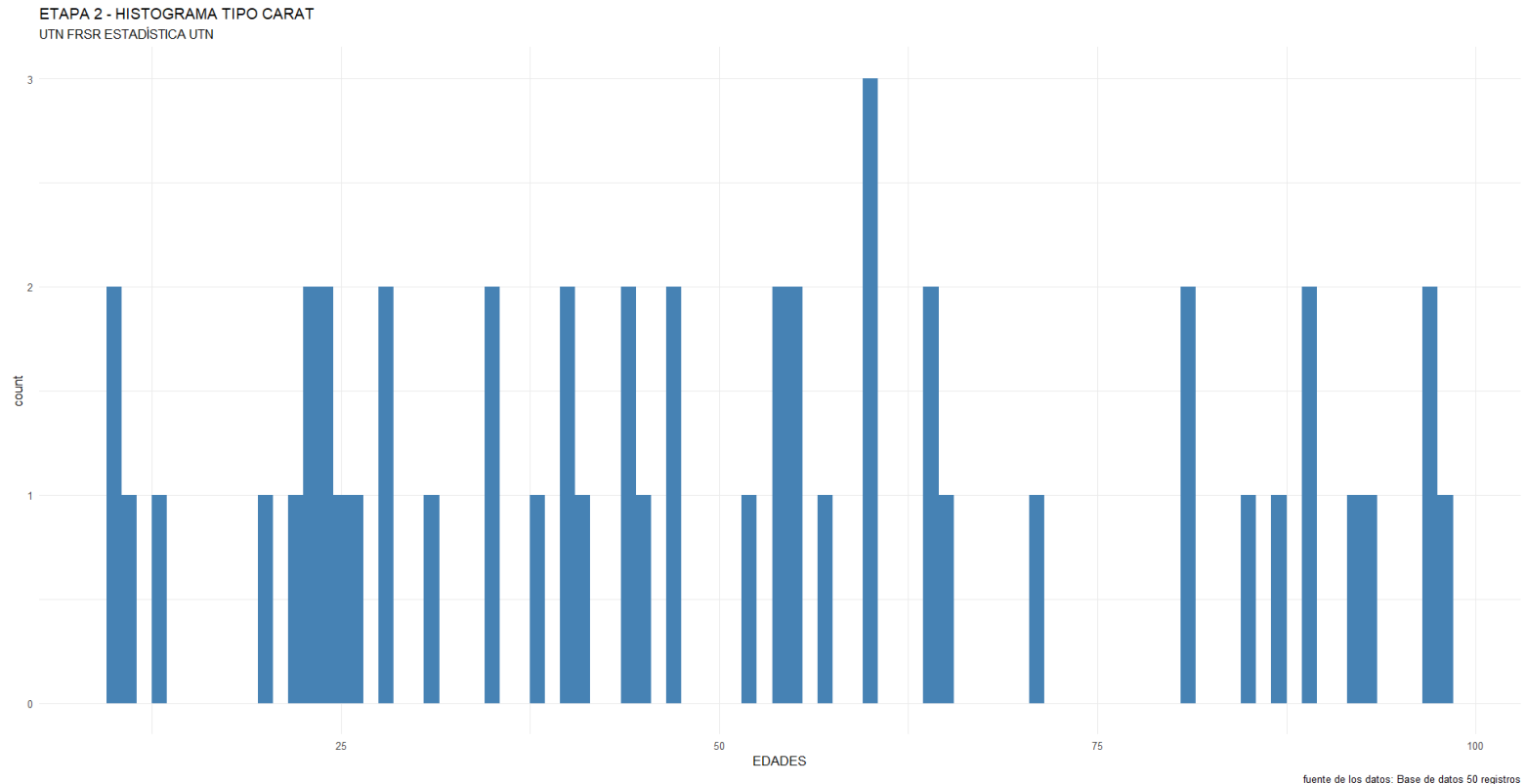
EL MISMO HISTOGRAMA PERO AHORA CON GGLOT2:

```
82  
83 library(ggplot2)  
84 # editando el histograma con ggplot2  
85 # ggplot2 se ve más lindo  
86 ggplot(data = baseDeDatos2,  
87        mapping = aes(x = EDAD)) +  
88   geom_histogram(bins = 9)  
89  
90 # separando por colores  
91 ggplot(data = baseDeDatos2,  
92        mapping = aes(x = EDAD,  
93                      fill = factor(NOMBRE))) +  
94   geom_histogram(bins = 9,  
95                 position = 'identity',  
96                 alpha = 0.8)+  
97   labs(title = 'ETAPA 2 - HISTOGRAMA',  
98        fill = 'Nombres',  
99        x = 'Edades',  
100       y = 'Conteo',  
101       subtitle = 'UTN FRSR ESTADÍSTICA UTN',  
102       caption = 'fuente de los datos: Base de datos 50 registros')  
103 )
```

Script del histograma anterior:

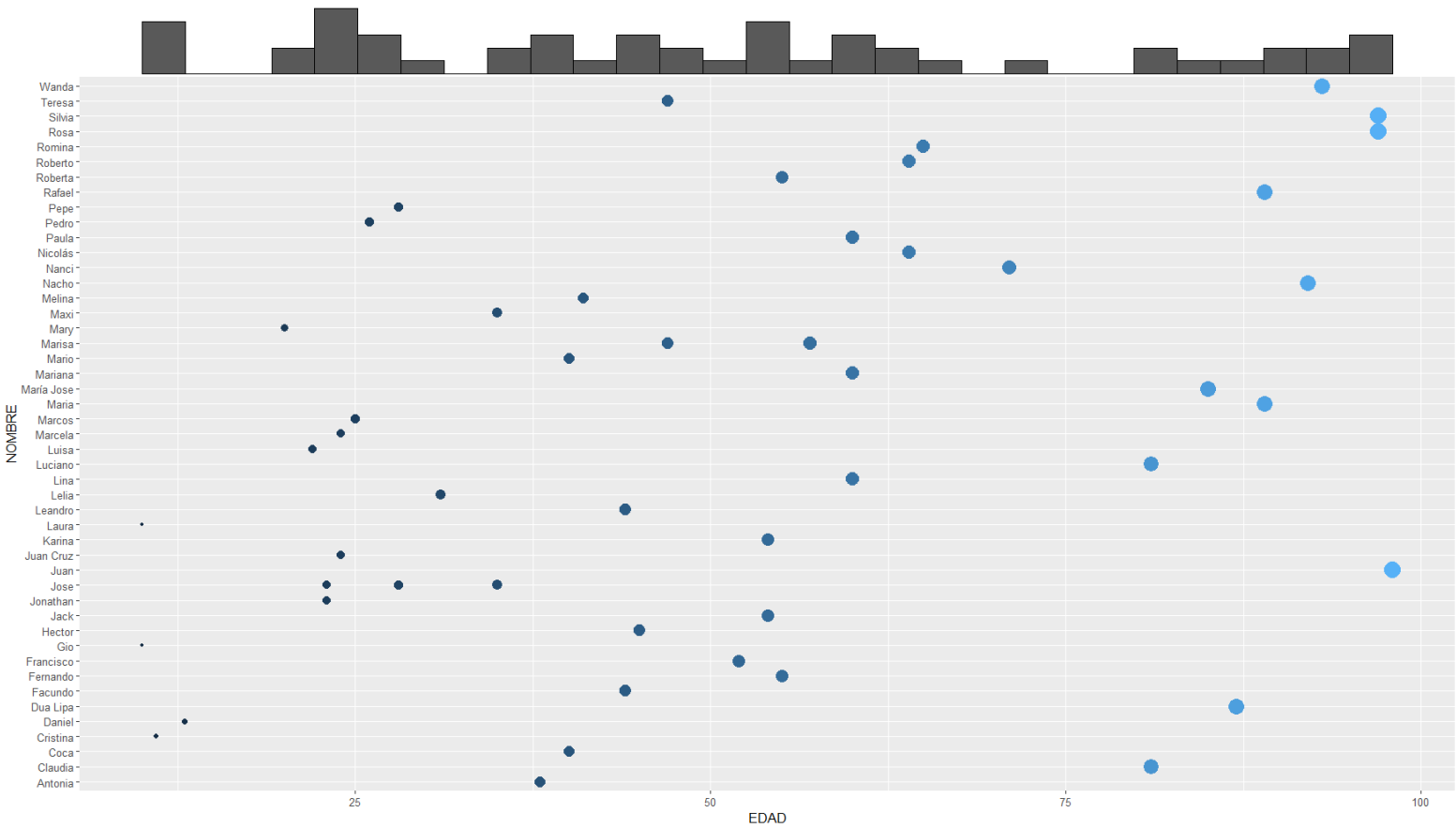


HISTOGRAMA TIPO CARAT: theme_minimal()



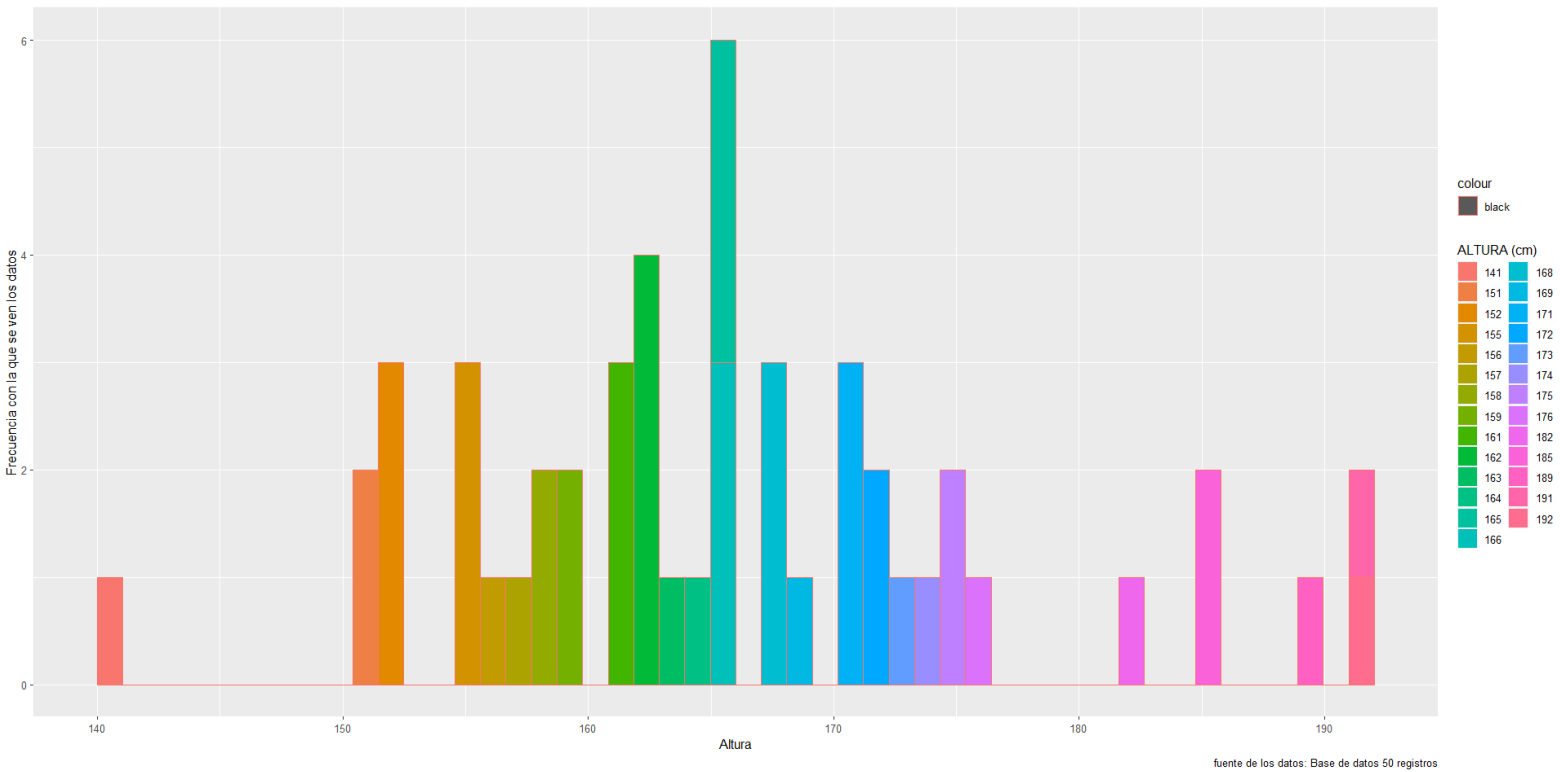
SCATTERPLOT: ggMarginal function (distribución marginal)

EJE X = edad EJE Y = nombres



HISTOGRAMA TIPO CARAT CON LA VARIABLE ALTURA

Histograma tipo CARAT con la variable altura
UTN FRSR ESTADÍSTICA UTN



Script del histograma anterior:

```

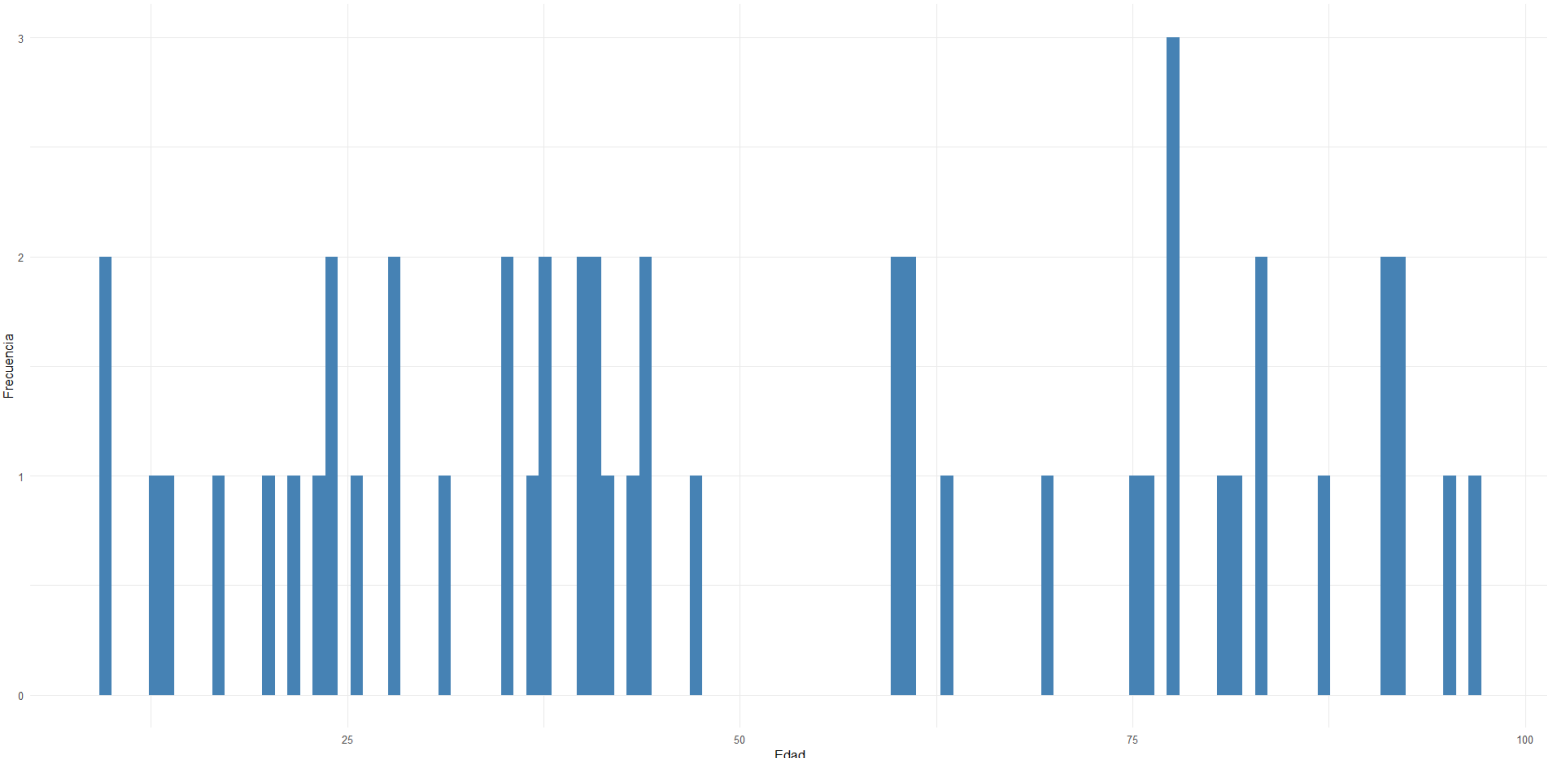
269 # histograma tipo CARAT en la variable ALTURA de nuestra database
270 ggplot(baseDeDatos2) +
271   geom_histogram(bins = 50, aes(x = ALTURA, fill = as.factor(ALTURA), color = 'black'))
272 + xlab("Altura") +
273   ylab("Frecuencia con la que se ven los datos") +
274   ggtitle("Histograma tipo CARAT con la variable altura")+
275   labs(fill="ALTURA (cm)", subtitle = "UTN FRSR ESTADÍSTICA UTN",
276        caption = "fuente de los datos: Base de datos 50 registros")
277 + theme_minimal()

```

HISTOGRAMA CON RANGO DE 0 A 2:

Distribución de la variable EDAD

intentando que cada columna tenga un rango de 0 a 2



Script del histograma con rango de 0 a 2:

```
282 # rango de 0 a 2 en histograma
283 ggplot(baseDeDatos2) +
284   geom_histogram(bins = 10, binwidth = 0.8, aes(x = EDAD), fill = 'steelblue') +
285   xlab("Edad") +
286   ylab("Frecuencia") +
287   ggtitle("Distribución de la variable EDAD",
288           subtitle = "intentando que cada columna tenga un rango de 0 a 2") +
289   theme_minimal()
290
291
```



HISTOGRAMA DE LAS VARIABLES PESO Y ALTURA CON CORTES:

Histograma de las variables Peso y Altura con cortes Cut

UTN FRSR ESTADÍSTICA

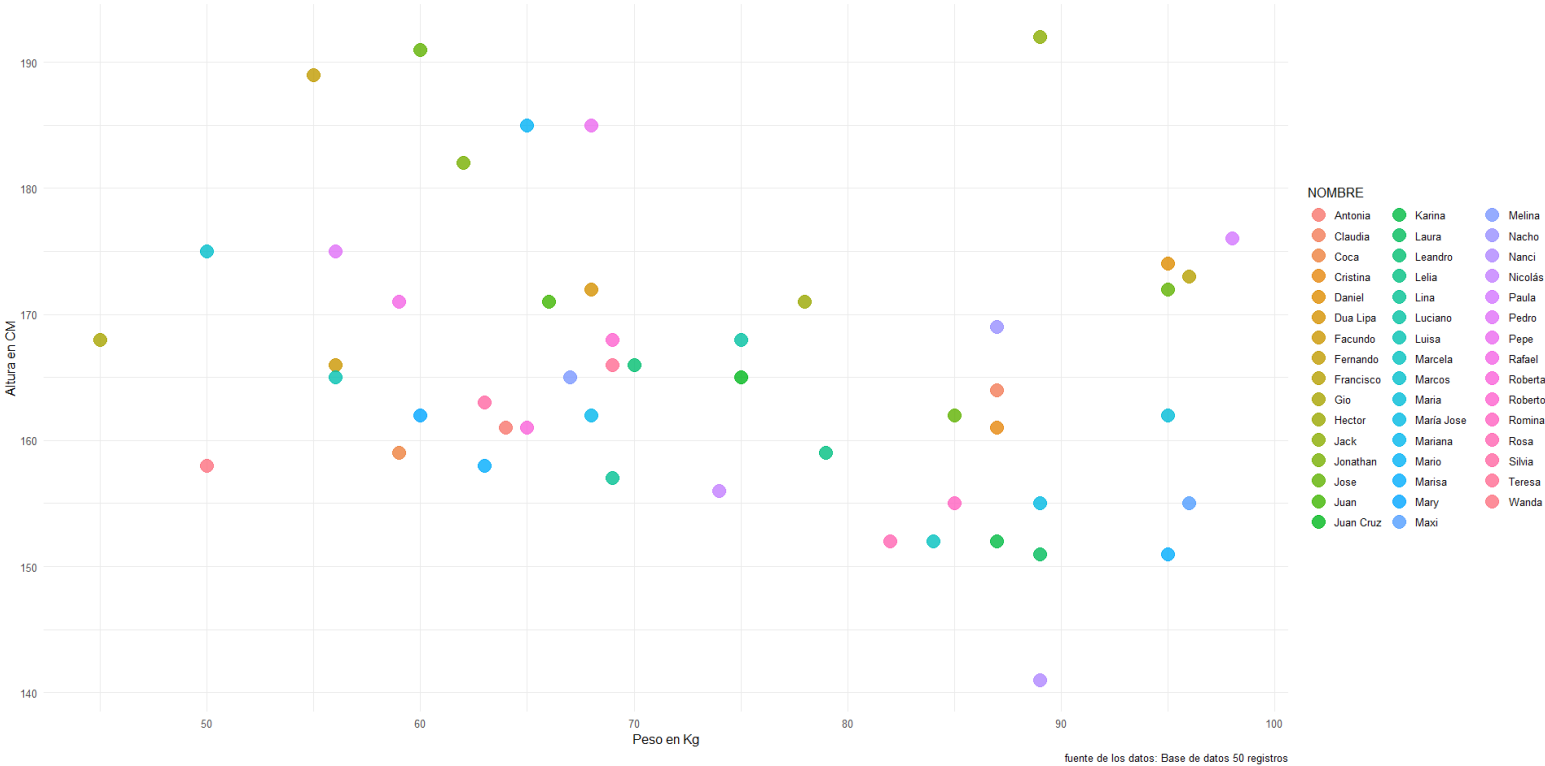


Script del histograma anterior:

```
280 # histograma de las variables Altura y Peso con cortes CUT
281 ggplot(baseDeDatos2) +
282   geom_histogram(bins = 50, aes(x = PESO, fill = as.factor(ALTURA))) +
283   facet_grid(ALTURA~., scales = 'free')+
284   xlab("Peso (Kg)") +
285   ylab("Frecuencia") +
286   ggtitle("Histograma de las variables Peso y Altura con cortes Cut",
287           subtitle = "UTN FRSR ESTADÍSTICA")+
288   labs(fill="Altura (cm)")+ theme_minimal()
289
```


GRÁFICO DE DISPERSIÓN Y DE PUNTOS:

Relación entre peso y edad de las personas
UTN FRSR ESTADÍSTICA

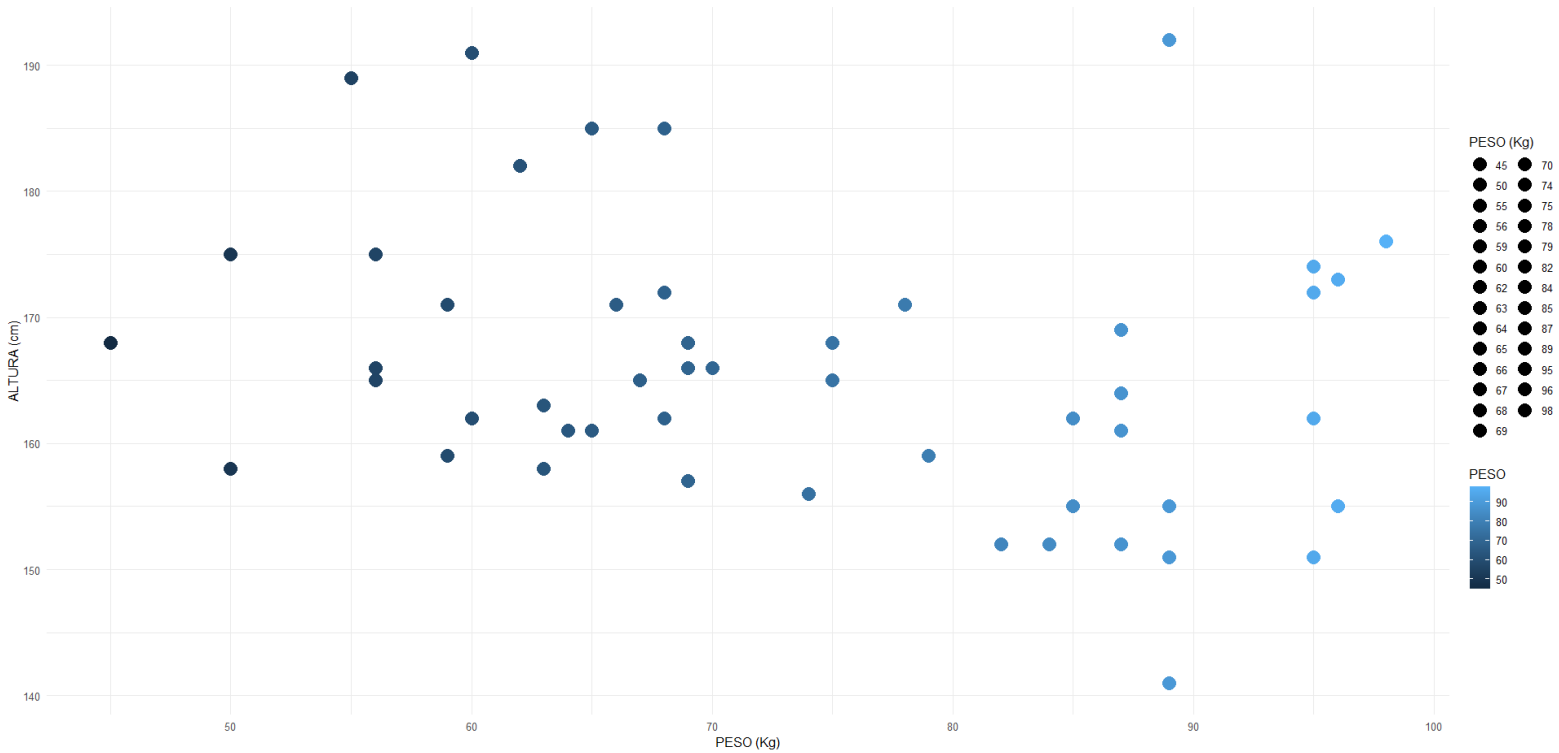


Script del gráfico:

```
344 # GRÁFICO DE DISPERSIÓN Y PUNTOS
345 ggplot(data = baseDeDatos2, aes(x = PESO, y = ALTURA)) +
346   geom_point(aes(color = NOMBRE), size = 5, alpha = 0.8) +
347   xlab('Peso en Kg') +
348   ylab('Altura en CM') +
349   ggtitle('Relación entre peso y edad de las personas',
350           subtitle = "UTN FRSR ESTADÍSTICA") +
351   labs(caption = "fuente de los datos: Base de datos 50 registros")+
352   theme_minimal()
353
```

GRÁFICO DE CAJAS BOXPLOT Y VIOLÍN

GRÁFICO DE CAJAS BOXPLOT Y VIOLÍN
UTN FRSR ESTADÍSTICA



Script del gráfico:

```
354 # GRÁFICO DE CAJAS BOXPLOT Y VIOLÍN
355 library("ggplot")
356 ggplot(data = baseDeDatos2,
357       aes(PESO,ALTURA))+
358   geom_point(aes(color = PESO,fill = as.factor(PESO)),
359 size=5,alpha=1)+
360   xlab('PESO (Kg)')+
361   ylab('ALTURA (cm)')+
362   ggtitle('GRÁFICO DE CAJAS BOXPLOT Y VIOLÍN',
363         subtitle = "UTN FRSR ESTADÍSTICA") +
364   labs(caption = "fuente de los datos: Base de datos 50 registros")+
365   theme_minimal()+
366   labs(fill="PESO (Kg)")
367 |
```