



Universidad del
Rosario

Facultad de
Economía

Laboratorio de
Finanzas

Escuela de Ingeniería,
Ciencia y Tecnología

TALLER DE **DATA SCIENCE EN R**



Sueño**SER**

Potencializa tu perfil
académico y profesional.

6:00 p.m a 8:00 p.m

10 al **12** de octubre

Trasciende,
construye
y Lidera

Rosarios
con
Propósito

Contenido

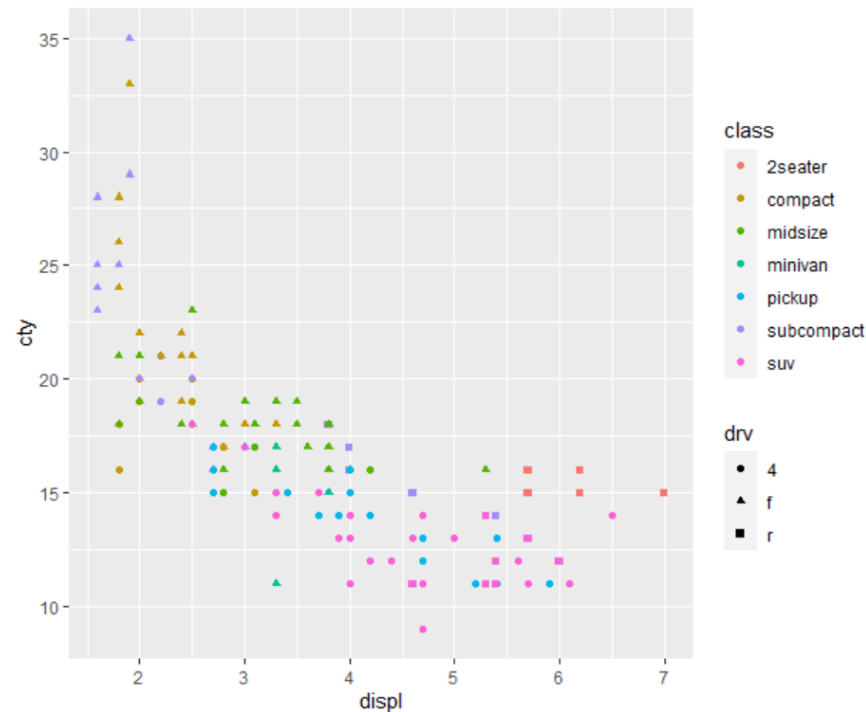
- Introducción
- Herramientas a emplear
- Gramática de los gráficos
- Primer gráfico
- Tipos de variables
- Gráficos más importantes para describir datos



Introducción

Algunas preguntas

- ¿Qué es un gráfico?
- ¿Qué tipos de gráficos existen?
- ¿Cómo describirías este gráfico?

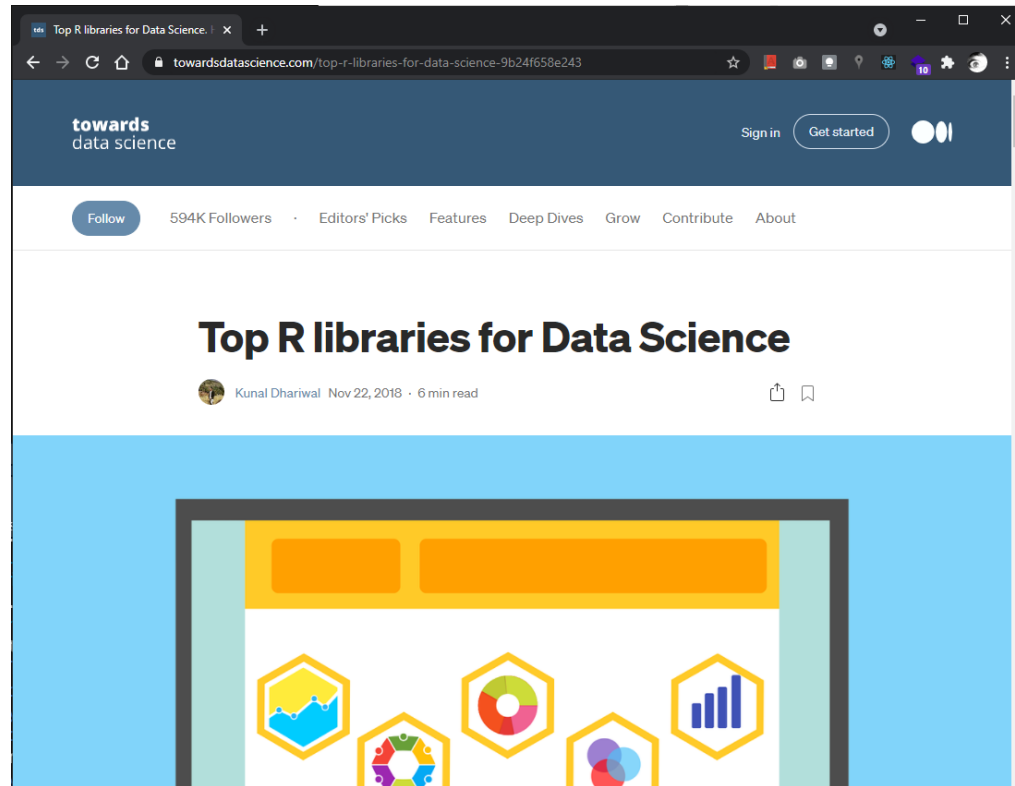




Herramientas a emplear

Instalación de paquetes

Herramientas a emplear



Se necesitan instalar una vez
por computador

```
install.packages("nombrepaquete")
```

pero hay que cargarlo cada vez
que se inicia R

```
library(nombredelpaquete)
```

Instalación de paquetes

Herramientas a emplear

```
# librerías importantes
```

```
library(readr)
```

```
library(tidyr)
```

```
library(dplyr)
```

```
library(ggplot2)
```



Paquete **tidyverse**

Herramientas a emplear

`library(tidyverse)` las incluye todas

```
> library(tidyverse)
-- Attaching packages ----- tidyverse 1.3.1 --
v ggplot2 3.3.3      v purrr   0.3.4
v tibble  3.1.1      v dplyr   1.0.5
v tidyr   1.1.3      v stringr 1.4.0
v readr   1.4.0      v forcats 0.5.1
-- Conflicts ----- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```



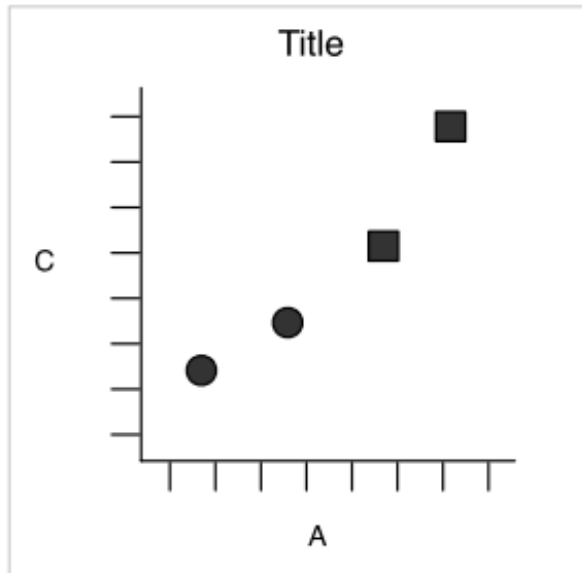

Gramática de los gráficos

Una gramática en capas

Componentes de un gráfico

- Datos y mapeos estéticos (**aesthetic mapping**)
 - Capa
 - Objetos geométricos
 - Transformación estadística
 - Ajustes de posición
 - Mapeos estéticos (opcionales)
 - Escalas
 - Sistema de **coordenadas**
 - Especificación de facetas (**facet**)
-
- <http://vita.had.co.nz/papers/layered-grammar.pdf>

Un dataset sencillo



Gramática de los gráficos

x	y	forma
25	11	circulo
0	0	circulo
75	53	cuadrado
200	300	cuadrado

Objetos gráficos producidos

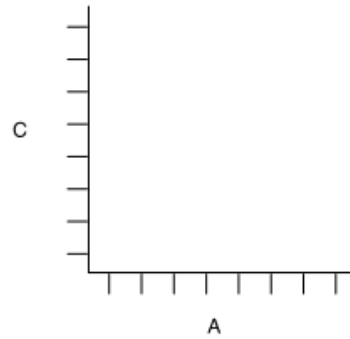
Gramática de los gráficos

Objetos
geométricos



+

Escala y
sistema de
coordenadas

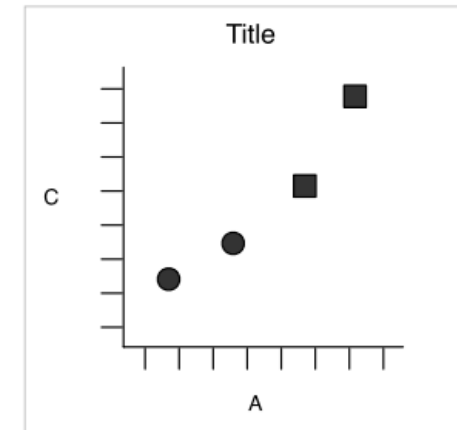


+

Anotaciones



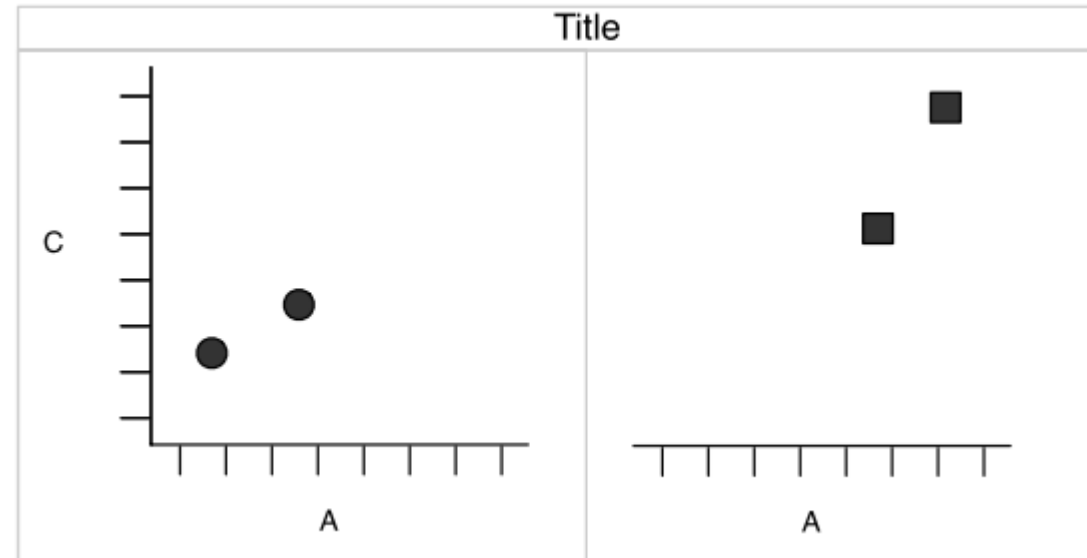
=



Facets

Gramática de los gráficos

x	y	forma
25	11	circulo
0	0	circulo
75	53	cuadrado
200	300	cuadrado



Un panel por cada tipo en la variable **forma**



Primer gráfico

Carga del dataset

```
pos_x <- c(25,0,75,200)
pos_y <- c(11,0,53,300)
forma <- c("circulo", "circulo", "cuadrado", "cuadrado")

datos <- data.frame(pos_x,pos_y,forma)

View(datos)
```

pos_x	pos_y	forma
25	11	circulo
0	0	circulo
75	53	cuadrado
200	300	cuadrado

Un template para cualquier gráfico

Primer gráfico

capa
nueva

```
ggplot(data = <DATA>) +  
  <GEOM_FUNCTION>(mapping = aes(<MAPPINGS>))
```

- **DATA**: los datos
- **GEOM_FUNCTION**: tipo de objetos geométricos (puntos, barras, densidad, boxplot,...)
- **MAPPINGS**: relaciona cada variable con su eje (x = var1, y = var2)

Creación de gráfico a partir de los datos

Primer gráfico

```
ggplot(data = datos) +  
  geom_point(aes(x = pos_x, y = pos_y))
```

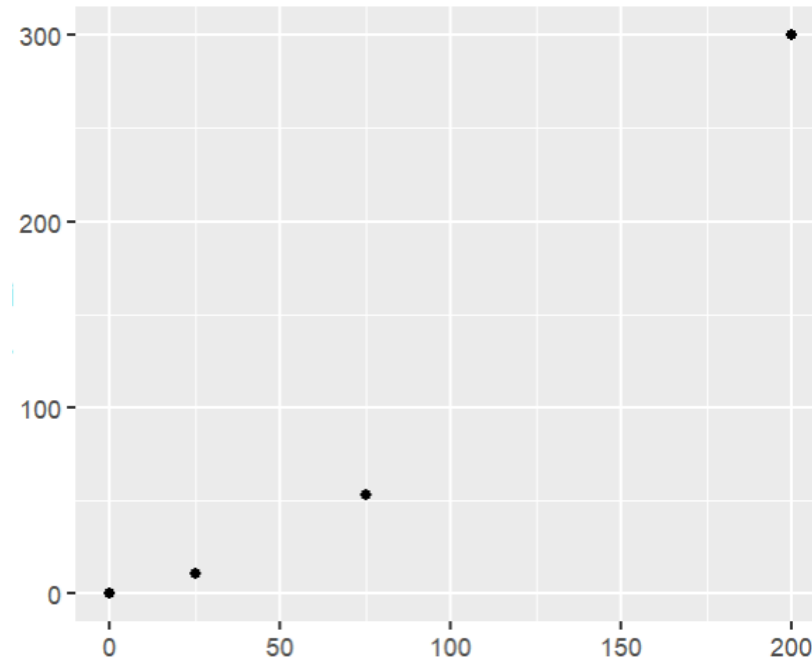


Gráfico de dispersión

Creación de gráfico a partir de los datos

Primer gráfico

```
ggplot(data = datos) +  
  geom_point(aes(x = pos_x, y = pos_y, shape = forma))
```

La estética incluye cosas como el tamaño (**size**), la forma (**shape**) o el color de los puntos (**color**). Puede mostrar un punto de diferentes formas cambiando los valores de sus propiedades estéticas.

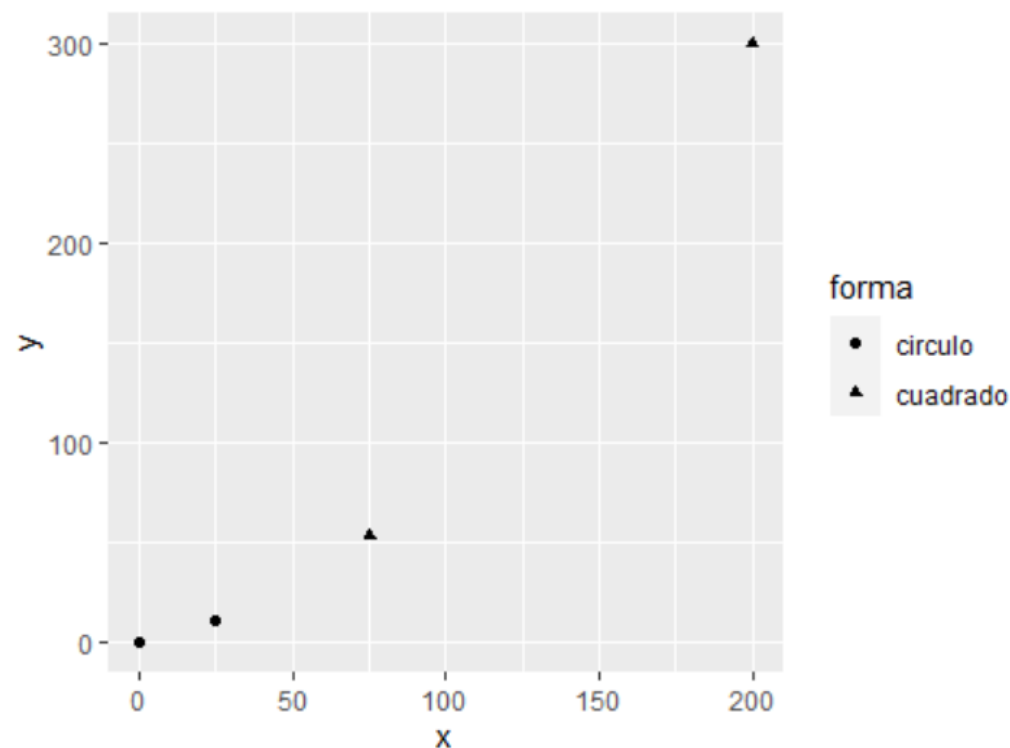


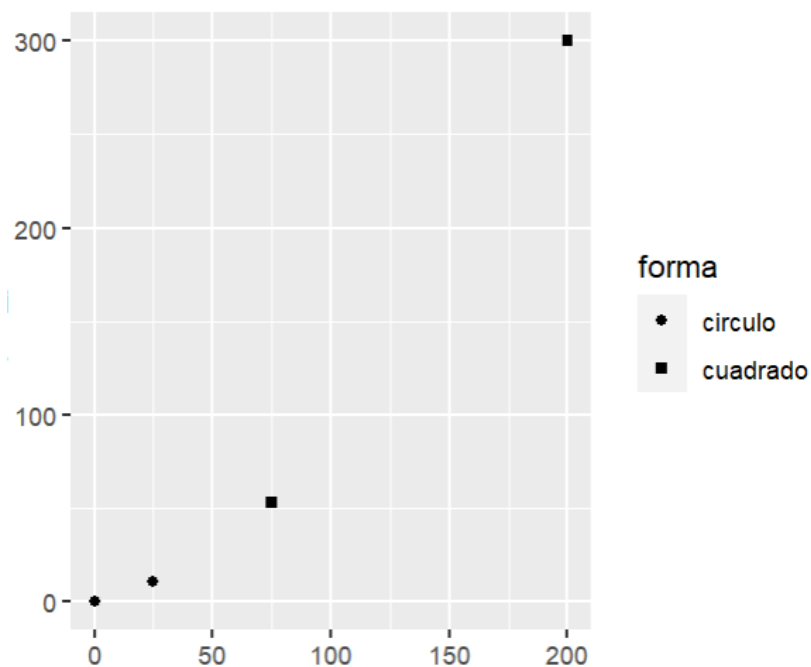
Gráfico de dispersión

Creación de gráfico a partir de los datos

Primer gráfico

```
ggplot(data = datos) +  
  geom_point(aes(x = pos_x, y = pos_y, shape = forma)) +  
  scale_shape_manual(values = c(16,15))
```

Quando se mapea una variable en `aes()` ggplot2 asigna automáticamente un nivel único de la estética (aquí un color único) a cada valor único de la variable, un proceso conocido como **scaling**



Creación de gráfico a partir de los datos

Primer gráfico

```
ggplot(data = datos) +  
  geom_point(aes(x = pos_x, y = pos_y, shape = forma), size = 5, color = "red") +  
  scale_shape_manual(values = c(16,15))
```

Se pueden asignar valores
de estéticas
manualmente,
ajustándolos por fuera del
`aes()`

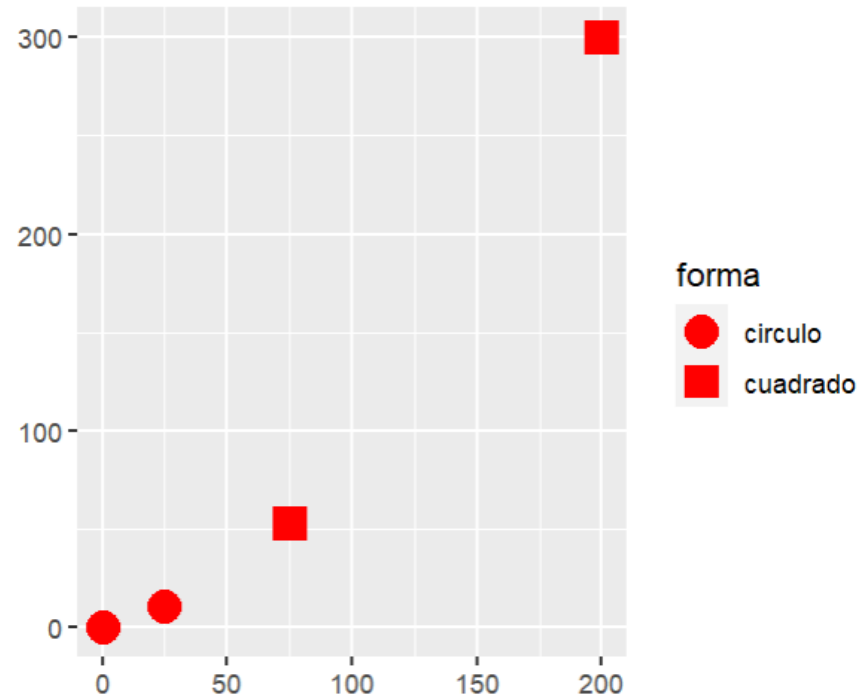


Gráfico de dispersión

El dataframe **mpg**

Primer gráfico

Contiene observaciones recolectadas por la EPA sobre 38 modelos de autos

Entre las variables tenemos

- displ: tamaño del motor en litros
- hwy: Eficiencia del combustible en carretera

ACTIVIDAD

Tu turno

1. Ejecuta `ggplot(data = mpg)`. ¿Qué ves?
2. ¿Cuántas filas y columnas tiene el dataset `mpg`?
3. ¿Qué quiere decir la variable `drv`? Revisa la ayuda
4. Elabora un gráficos de dispersion entre `hwy` y `cyl`.
5. ¿Qué pasa si hacemos un gráficos de dispersion de `class` y `drv`? ¿Por qué no es de utilidad el gráfico?

El dataframe **mpg**

Primer gráfico

Contiene observaciones recolectadas por la EPA sobre 38 modelos de autos

Entre las variables tenemos

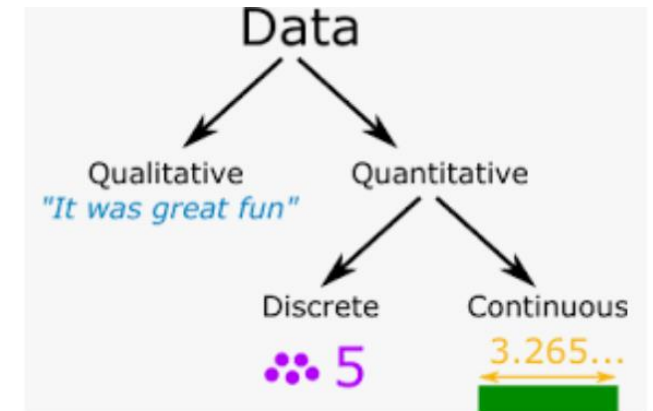
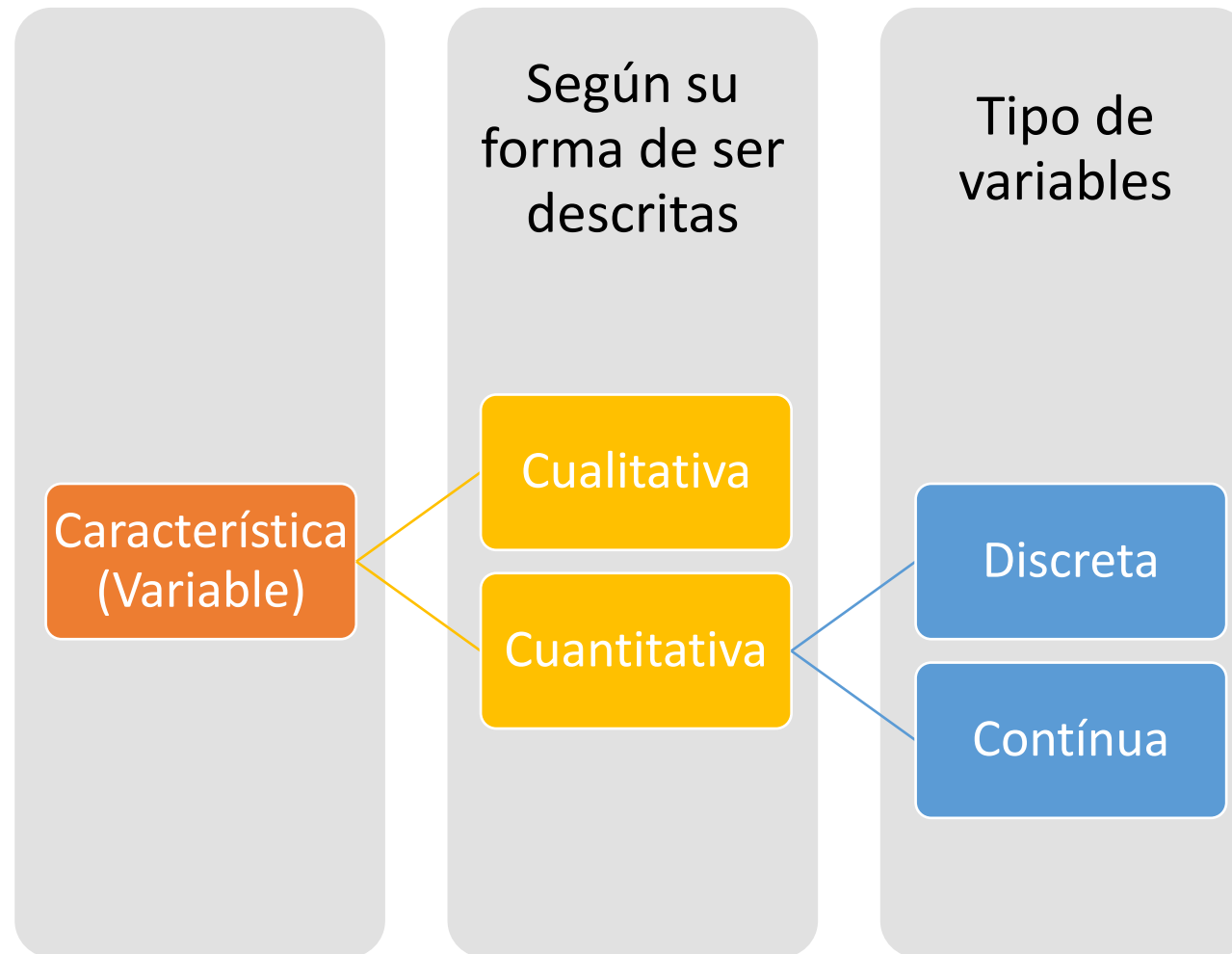
- displ: tamaño del motor en litros
- hwy: Eficiencia del combustible en carretera



Tipos de variables

Tipos de variables

Tipos de variables

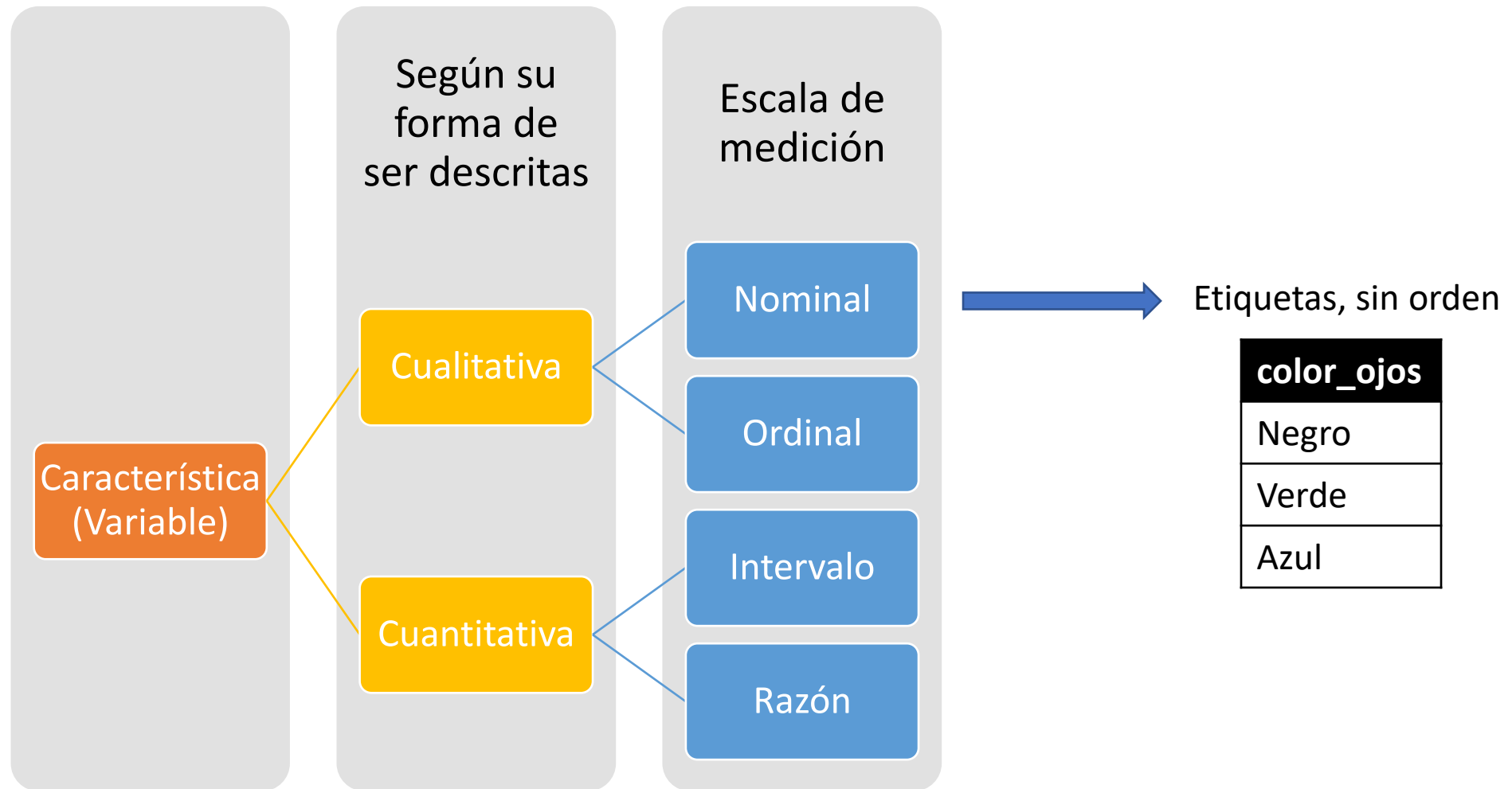


→ **Números enteros**
No valores intermedios
p.ej. núm. de mascotas

→ **Admite valores fraccionarios**
p.ej. estatura

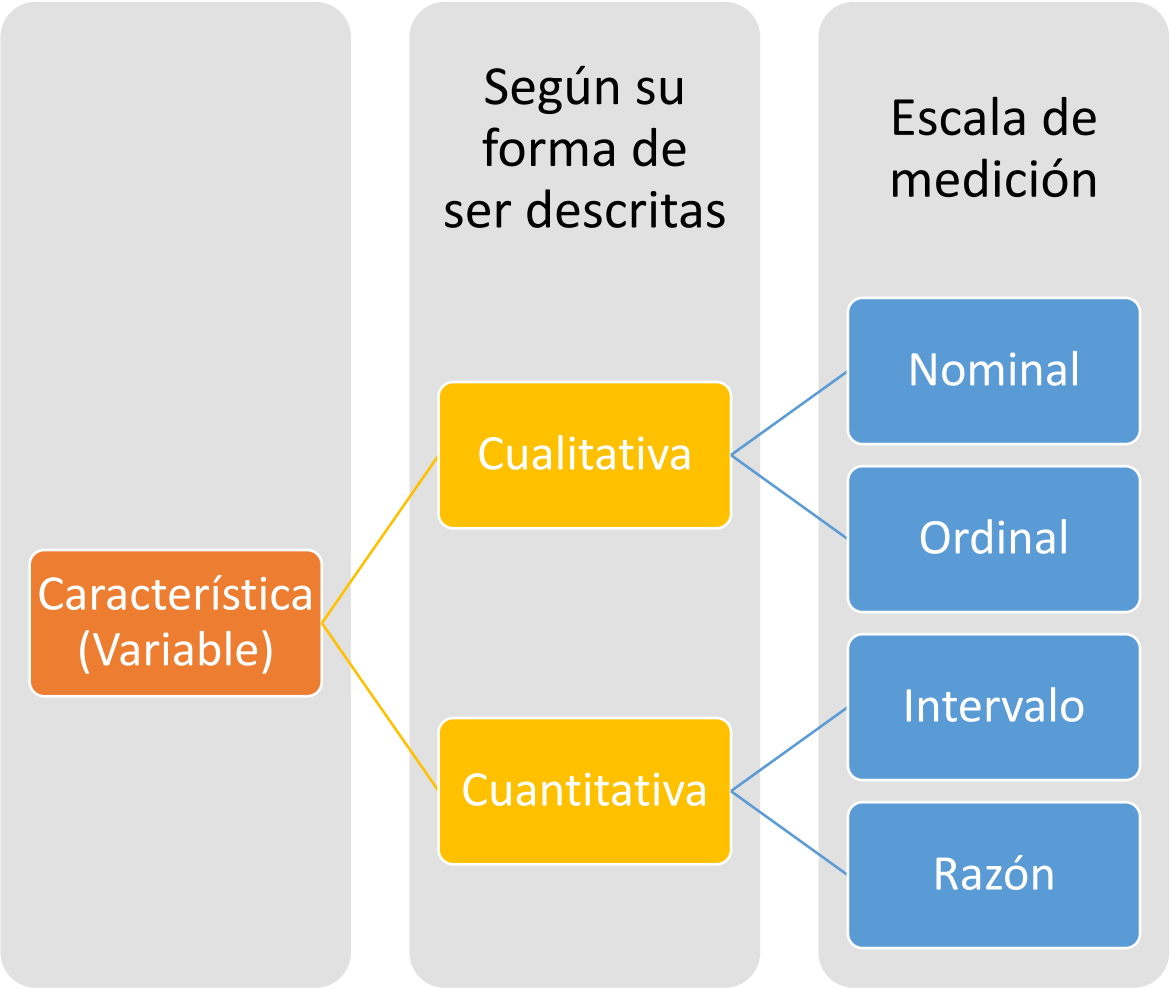
Escalas de medición

Tipos de variables



Escalas de medición

Tipos de variables

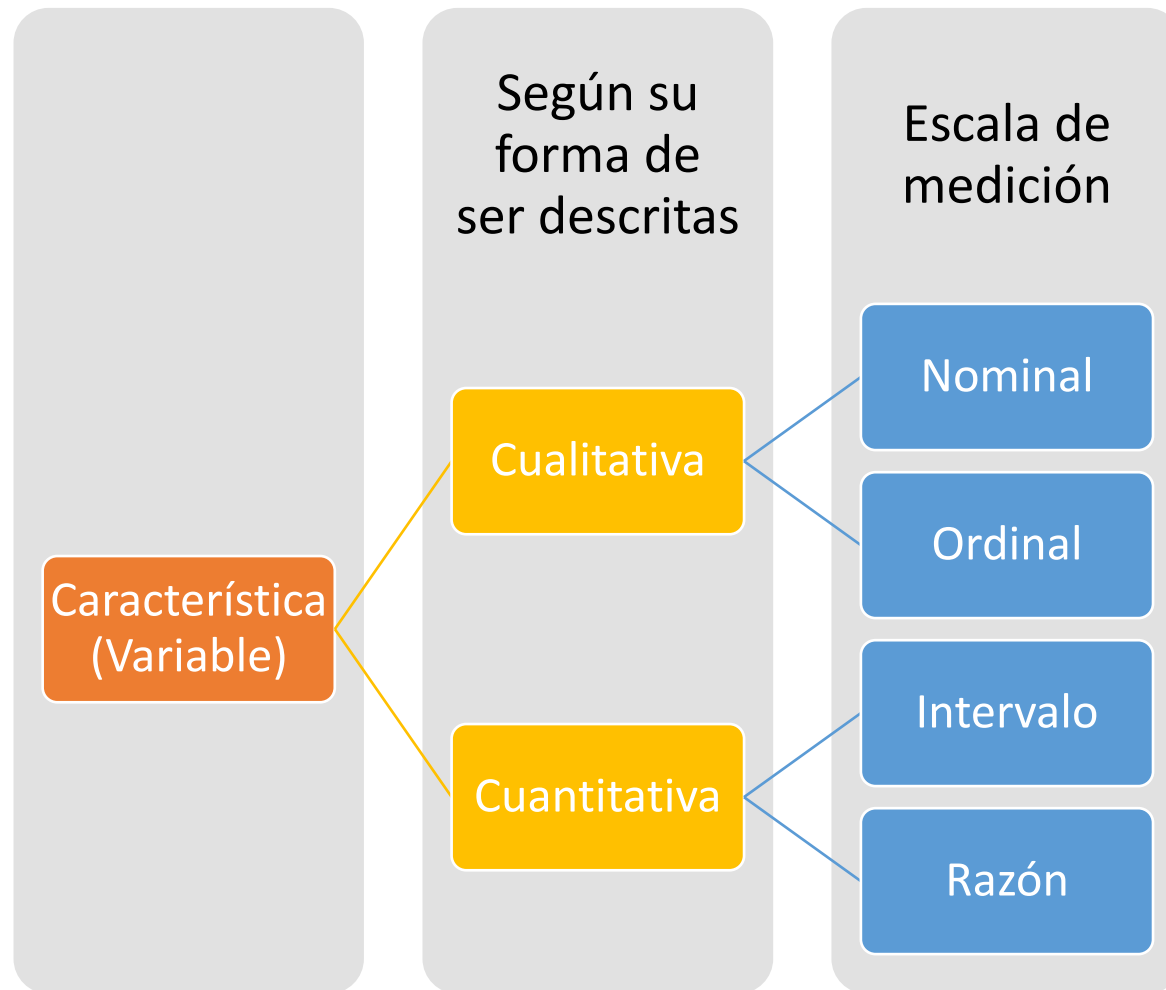


Etiquetas, con orden

estado_salud
Muy saludable
Medianamente saludable
Saludable
Poco saludable
No saludable

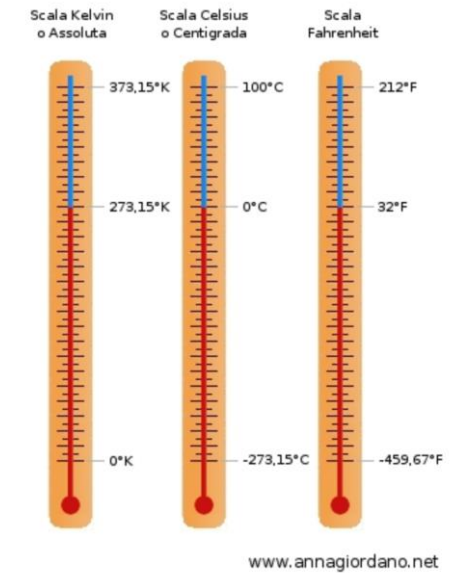
nivel_socioeconómico
Alto
Medio
Bajo

Escalas de medición



temp_c
36.8
38.4
37.1

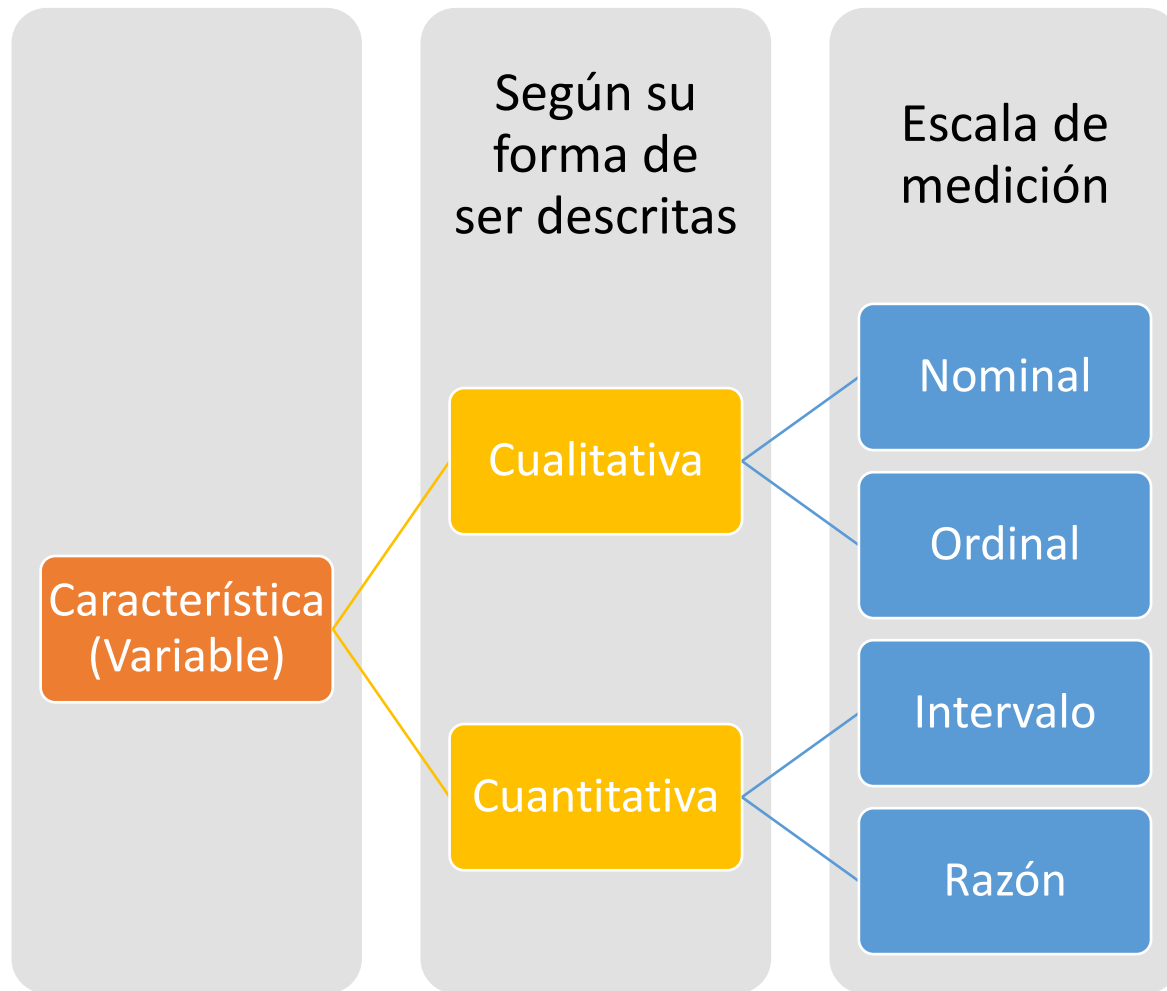
Tipos de variables



Valores numéricos.
Unidades de medidas constantes.
Se conoce la distancia exacta entre mediciones.
Se pueden hacer comparaciones entre la diferencia de mediciones.
El cero es una **convención**.

Escalas de medición

Tipos de variables



precio
\$5000
\$3200
\$0

Valores numéricos.
Unidades de medidas constantes.
Se conoce la distancia exacta entre mediciones.
Se pueden hacer comparaciones entre la diferencia de mediciones.
El cero es **absoluto**, representa **ausencia** de lo que se mide.

Escala de medición y sus operaciones posibles

Tipos de variables

Nivel de medición	Identificación: Hay distinción entre categorías	Orden: Se pueden ordenar	Unidad de medida constante: Se conoce la distancia exacta entre cada categoría	Cero absoluto: Ausencia de valor en la escala que se traduzca
Operaciones	Contar	Ordenar	Comparar diferencias	Comparar razones
Relaciones posibles	$=, \neq$	$<, >$	$+, -$	\times, \div
Nominal	✓			
Ordinal	✓	✓		
Intervalo	✓	✓	✓	
Razón	✓	✓	✓	✓

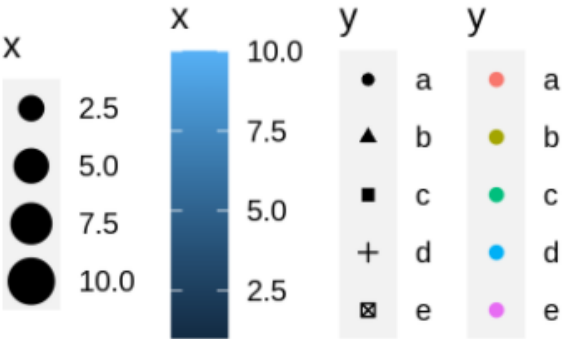
Tipos de datos en R

Tipos de variables

The screenshot shows the RStudio Environment pane for a project named 'Fwd_clase_4'. The 'Data' tab is selected, showing a data frame named 'datos' with 4 observations and 3 variables. The variables are 'x' (numeric), 'y' (numeric), and 'forma' (character). The 'Values' tab shows the data for each variable. The 'forma' variable is highlighted with an orange box, and an orange arrow points to it from the left. Another orange arrow points to the 'Data' tab. The 'Values' tab shows the following data:

Variable	Forma	x	y
datos	"circulo"	25	11
	"circulo"	0	0
	"cuadrado"	75	53
	"cuadrado"	200	300

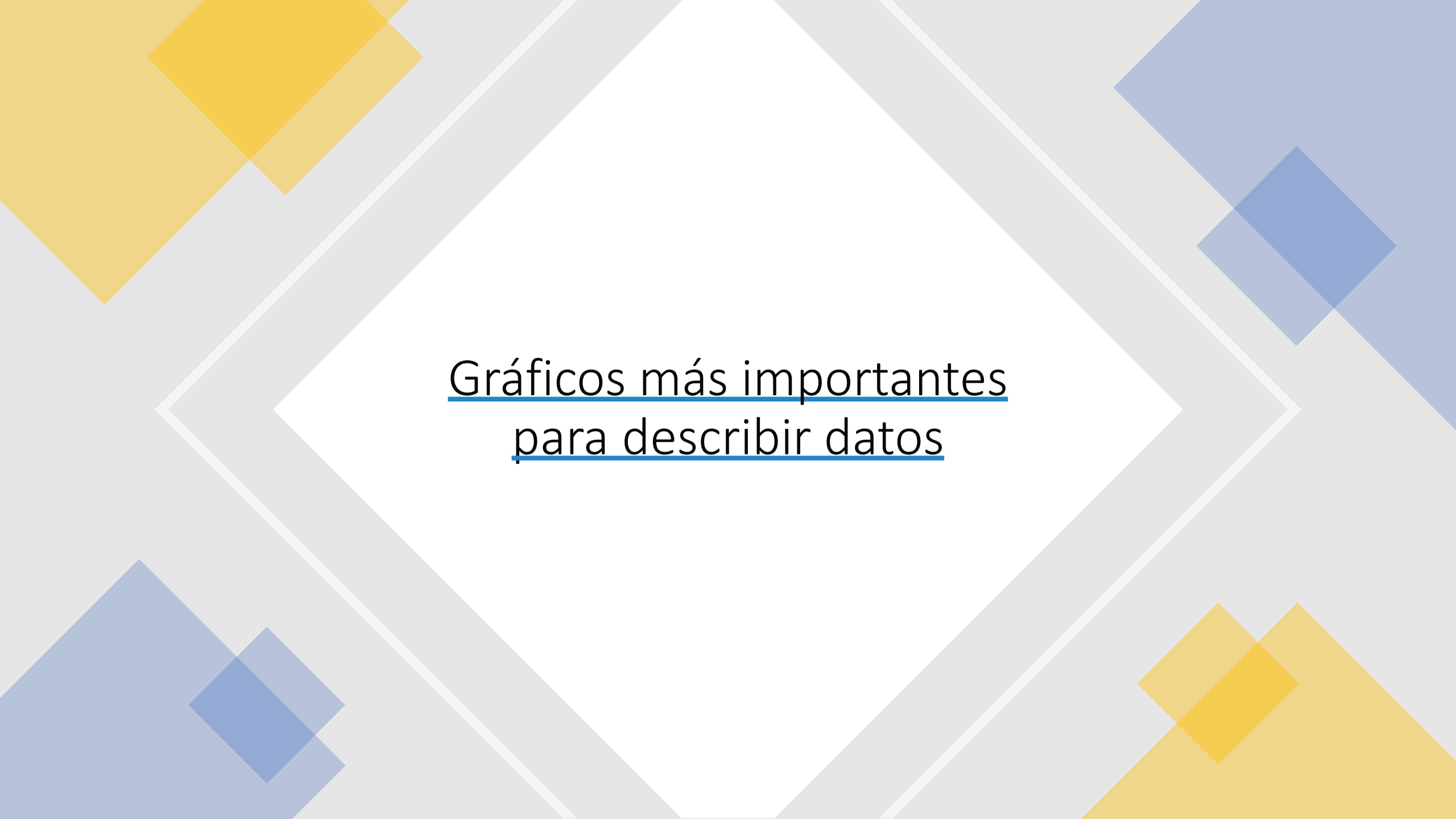
Abreviación	Significado
int	Número entero
num,	Número real
chr	Cadena de caracteres
logi	Booleanos
POSIXct	Fecha/Hora
Factor	Categorías



Algunas transformaciones estadísticas suministradas por ggplot2

Gráficos más importantes para
describir datos

Name	Description
bin	Divide continuous range into bins, and count number of points in each
boxplot	Compute statistics necessary for boxplot
contour	Calculate contour lines
density	Compute 1d density estimate
identity	Identity transformation, $f(x) = x$
jitter	Jitter values by adding small random value
qq	Calculate values for quantile-quantile plot
quantile	Quantile regression
smooth	Smoothed conditional mean of y given x
summary	Aggregate values of y for given x
unique	Remove duplicated observations



Gráficos más importantes para describir datos

De una variable

Gráficos más importantes para describir datos

ONE VARIABLE continuous

```
c <- ggplot(mpg, aes(hwy)); c2 <- ggplot(mpg)
```



c + geom_area(stat = "bin")

x, y, alpha, color, fill, linetype, size



c + geom_density(kernel = "gaussian")

x, y, alpha, color, fill, group, linetype, size, weight



c + geom_dotplot()

x, y, alpha, color, fill



c + geom_freqpoly() x, y, alpha, color, group,

linetype, size



c + geom_histogram(binwidth = 5) x, y, alpha,

color, fill, linetype, size, weight



c2 + geom_qq(aes(sample = hwy)) x, y, alpha,

color, fill, linetype, size, weight

discrete

```
d <- ggplot(mpg, aes(fl))
```



d + geom_bar()

x, alpha, color, fill, linetype, size, weight

<https://www.maths.usyd.edu.au/u/UG/SM/STAT3022/r/current/Misc/data-visualization-2.1.pdf>

Dos variables

Gráficos más importantes para describir datos

TWO VARIABLES

continuous x , continuous y

```
e <- ggplot(mpg, aes(cty, hwy))
```



e + geom_label() (aes(label = cty), nudge_x = 1, nudge_y = 1, check_overlap = TRUE) x, y, label, alpha, angle, color, family, fontface, hjust, lineheight, size, vjust



e + geom_jitter() (height = 2, width = 2) x, y, alpha, color, fill, shape, size



e + geom_point(), x, y, alpha, color, fill, shape, size, stroke



e + geom_quantile(), x, y, alpha, color, group, linetype, size, weight



e + geom_rug() (sides = "bl"), x, y, alpha, color, linetype, size



e + geom_smooth() (method = lm), x, y, alpha, color, fill, group, linetype, size, weight



e + geom_text() (aes(label = cty), nudge_x = 1, nudge_y = 1, check_overlap = TRUE), x, y, label, alpha, angle, color, family, fontface, hjust, lineheight, size, vjust

discrete x , discrete y

```
g <- ggplot(diamonds, aes(cut, color))
```



g + geom_count(), x, y, alpha, color, fill, shape, size, stroke

discrete x , continuous y

```
f <- ggplot(mpg, aes(class, hwy))
```



f + geom_col(), x, y, alpha, color, fill, group, linetype, size



f + geom_boxplot(), x, y, lower, middle, upper, ymax, ymin, alpha, color, fill, group, linetype, shape, size, weight



f + geom_dotplot() (binaxis = "y", stackdir = "center"), x, y, alpha, color, fill, group



f + geom_violin() (scale = "area"), x, y, alpha, color, fill, group, linetype, size, weight

continuous function

```
i <- ggplot(economics, aes(date, unemploy))
```



i + geom_area(), x, y, alpha, color, fill, linetype, size



i + geom_line(), x, y, alpha, color, group, linetype, size



i + geom_step() (direction = "hv"), x, y, alpha, color, group, linetype, size

Gracias por tu asistencia y participación 😊

Contacto

✉ miguela.orjuela@urosario.edu.co

🌐 <https://www.linkedin.com/in/miguel-orjuela/>

🐙 <https://github.com/maorjuela73>

Links de interés

- <https://ggplot2-book.org/>
- <https://r4ds.had.co.nz/>
- <https://raw.githubusercontent.com/rstudio/cheatsheets/main/data-visualization.pdf>