



## (2) Instalación de R y RStudio. Importación y visualización de datos

Rubén Heradio  
Universidad Nacional de Educación a Distancia

Plasencia 2025 – Curso de Análisis y Visualización de Datos: Estadística  
Práctica con R e Inteligencia Artificial

# Instalación



[https://github.com/  
rheradio/curso\\_  
verano\\_analisis\\_datos](https://github.com/rheradio/curso_verano_analisis_datos)

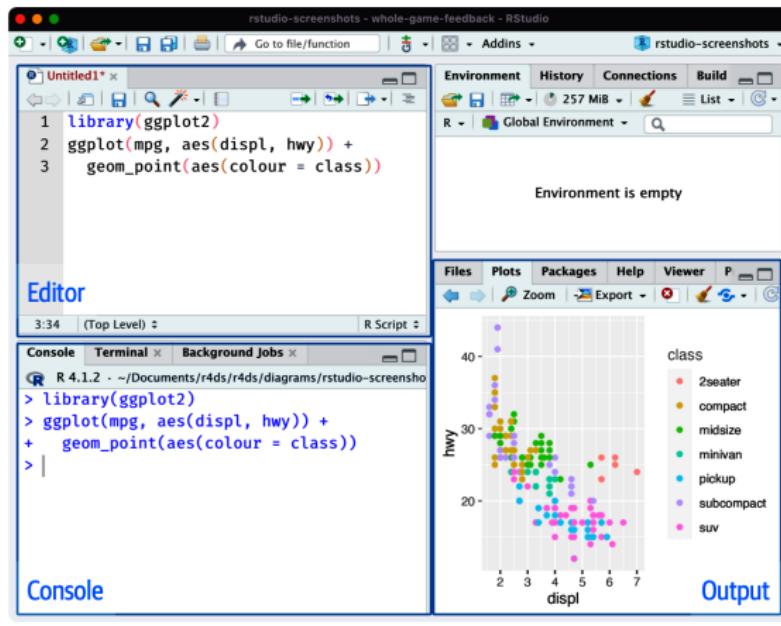
① R:

<https://cran.r-project.org/>

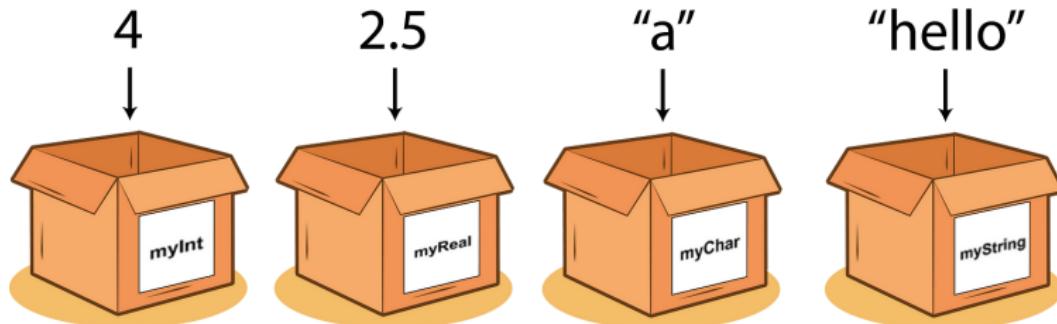
② RStudio:

[https://posit.co/products/  
open-source/rstudio/](https://posit.co/products/open-source/rstudio/)

# RStudio



# Variables



```
1 mi_numero <- 8.28
2 mi_texto <- "hola"
```

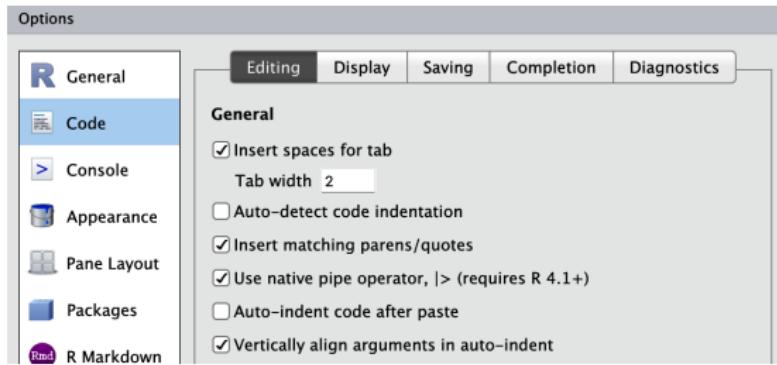
Atajo para escribir <- en windows: Alt + -

# Funciones

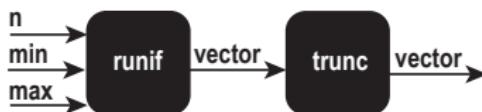


```
1 runif(n=5, min=0, max=10)
2 runif(min=0, n=5, max=10)
3 runif(5, 0, 10)
```

# Pipe



# Encadenando funciones

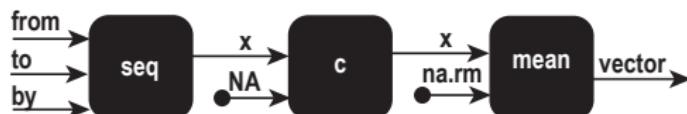


```
1  trunc(runif(5, 0, 10))
```

```
1  x <- runif(5, 0, 10)
2  trunc(x)
```

```
1  runif(5, 0, 10) |>
2  trunc()
```

# Ejercicio



1	0.00	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00
2	3.25	3.50	3.75	4.00	4.25	4.50	4.75	5.00	5.25	5.50	5.75	6.00	6.25
3	6.50	6.75	7.00	7.25	7.50	7.75	8.00	8.25	8.50	8.75	9.00	9.25	9.50
4	9.75	10.00											

1	0.00	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00
2	3.25	3.50	3.75	4.00	4.25	4.50	4.75	5.00	5.25	5.50	5.75	6.00	6.25
3	6.50	6.75	7.00	7.25	7.50	7.75	8.00	8.25	8.50	8.75	9.00	9.25	9.50
4	9.75	10.00	NA										

1	5
---	---

# Paquetes de funciones

- En [https://cran.r-project.org/web/packages/available\\_packages\\_by\\_name.html](https://cran.r-project.org/web/packages/available_packages_by_name.html) hay **22.448 paquetes**
- Un paquete se **instala** con Tools > Install Packages...
- Un paquete se **importa** escribiendo `library(paquete)`

# Tidyverse

Vamos a instalar el paquete más importante del curso: tidyverse



- Español (edición 1):  
<https://es.r4ds.hadley.nz/>
- Inglés (edición 2):  
<https://r4ds.hadley.nz/>

# Rango

En 1880, Francis Galton estudió la relación entre la estatura de padres e hijos. Vamos a centrarnos en la de los hijos.

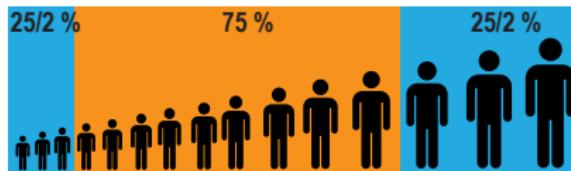
¿Qué medidas describen mejor los valores muestrales y su variación?

- min: 1.42 m
- max: 2.01 m

El rango es demasiado sensible a los **valores extremos**. Mejor pensar en “la variación típica” que en la “variación extrema”

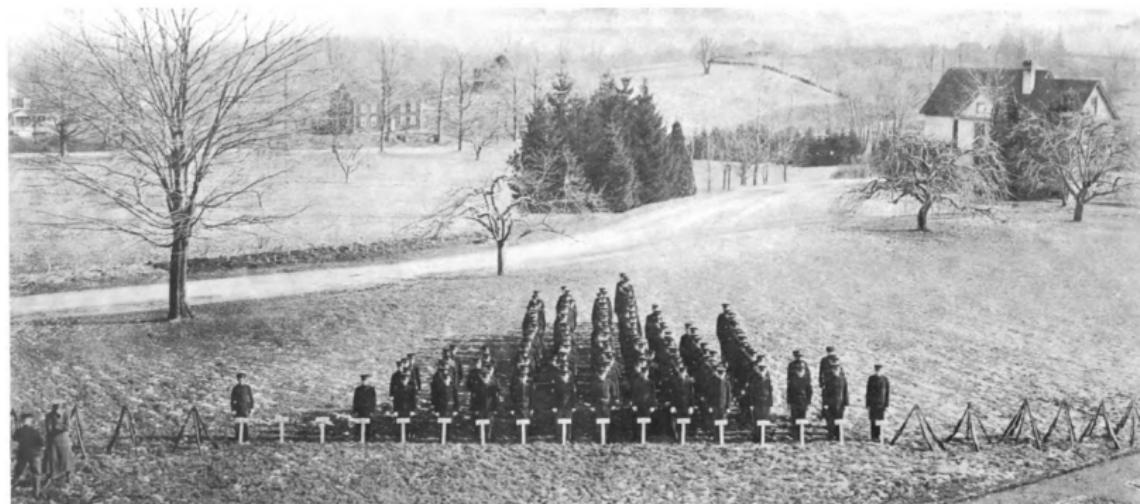


# Intervalo de valores centrales



- Para protegernos de que la muestra tenga valores atípicos, flexibilizamos los extremos
- Ordenamos el conjunto de valores de menor a mayor
- El percentil  $n$  es el valor que divide al conjunto de datos de tal forma que el  $n\%$  de los datos son menores que dicho valor
- La mediana es el percentil 50 %
- No confundir **coverage interval** con **confidence interval**

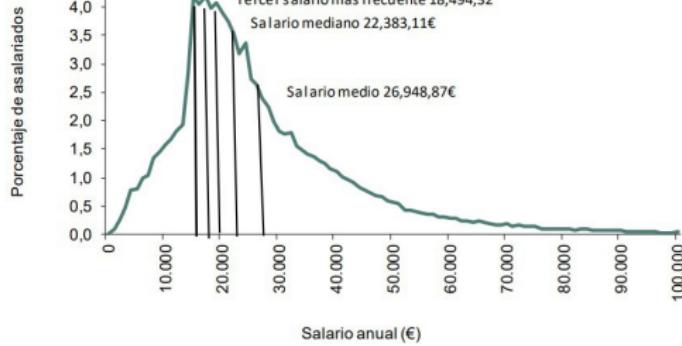
# Histograma



Number of individuals in each rank 1 0 0 1 5 7 7 22 25 26 27 17 11 17 4 4 1  
Heights in feet and inches to which  
ranks correspond . . . . . 4:10 4:11 5.0 5:1 5:2 5:3 5:4 5:5 5:6 5:7 5:8 5:9 5:11 6:0 6:1 6:2

# Media y varianza

- Media:  $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$
- Varianza:  $\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$
- Desviación típica:  $\sigma = \sqrt{\sigma^2}$



- Facilitan las matemáticas
- Son inútiles cuando las distribuciones son asimétricas
- Muy sensibles a valores extremos

# ggplot2

- ggplot2 proporciona un lenguaje uniforme para describir gráficas
- Chuleta: [https://github.com/rheradio/curso\\_verano\\_analisis\\_datos/blob/main/transparencias/chuleta\\_ggplot2.pdf](https://github.com/rheradio/curso_verano_analisis_datos/blob/main/transparencias/chuleta_ggplot2.pdf)
- La última figura visualizada puede guardarse con “ggsave”. Si la figura es sencilla, usa PDF. Si tiene mucha información, usa PNG.

# Número de parejas a lo largo de la vida

En los 80, el SIDA era un problema enorme. En el Reino Unido, se realizó un estudio sobre las relaciones sexuales con personas diferentes que tenían sus ciudadanos a lo largo de su vida.

Vamos a aprender a pintar histogramas y gráficas de densidad:

- Los histogramas son muy sensibles al número de *bins*
- Mejor usar gráficas de densidad

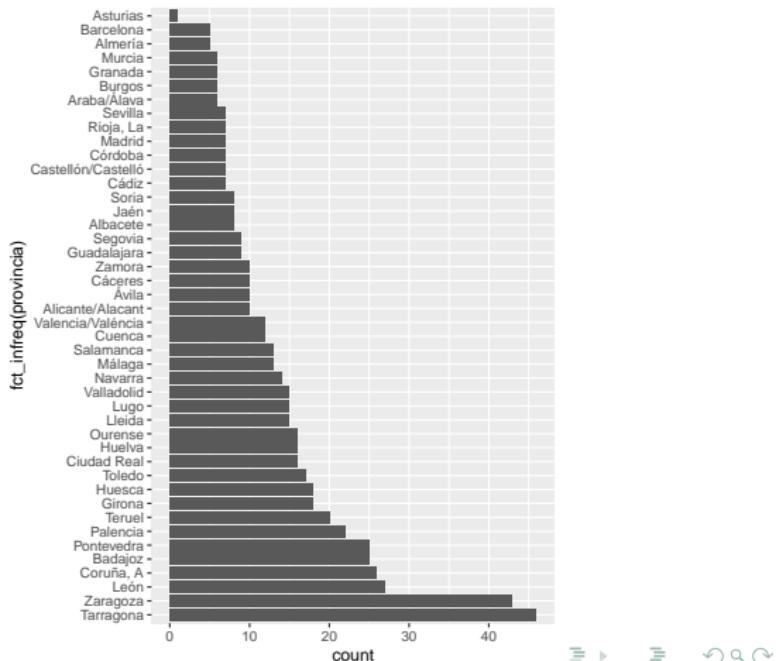
# Número de pelis de James Bond por cada actor

Vamos a aprender a pintar gráficas de barras y diagramas de tarta:

- Las gráficas de barras son el equivalente categórico a los histogramas
- Los diagramas de tarta sólo son útiles cuando hay pocas categorías, que además tienen una frecuencia muy diferente entre sí.

# Ejercicio

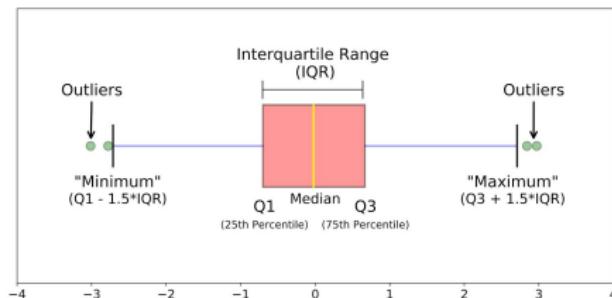
Utilizando  
“datos/estaciones\_tren\_larga  
\_y\_media\_distancia.csv” ,  
visualiza el  
número de  
estaciones por  
provincia



## Numérica - Categórica

Éste es el caso típico de los t-tests y ANOVA.

- Los histogramas y las gráficas de densidad pierden utilidad, porque suelen solaparse
- Lo mejor es usar varios diagrama de caja (uno por categoría).



- ggplot2 incluye un buen número de geoms. Además de éstos, hay muchas extensiones disponibles en:  
<https://exts.ggplot2.tidyverse.org/gallery/>

## Ejercicio

¿Hay una correlación negativa entre el dinero que cuesta una peli de James Bond y su puntuación en IMDB?

- Haz un diagrama de dispersión con el presupuesto de cada peli en el eje x, y la puntuación en el eje y
- Colorea los puntos según el año. ¿Se ve algún patrón?
- Usa la función “if\_else” para convertir la variable Year en categórica, valiendo TRUE si la peli se ha estrenado en los últimos 25 años.

```
1 james_bond |>  
2   mutate(last_25_years=if_else(Year<2000 , FALSE , TRUE))
```

- Colorea los puntos según esta nueva variable. Ahora, ¿se aprecia algún patrón?

# Overplotting y ordenación de variables categóricas

- Cuando en los diagramas de dispersión hay muchos datos, se suele producir el solapamiento entre los puntos. Para mejorar la visualización, utiliza “alpha” y “facets”
- En R, las variables categóricas se denominan “factors” y se pueden ordenar según su propiedad “levels”

# Etiquetas y diversas fuentes de datos

- Usa “geom\_text\_repel” para etiquetar puntos
- Podemos pintar columnas nuevas obtenidas sobre la marcha combinando otras columnas existentes
- Cada geom puede tener una fuente de datos independiente

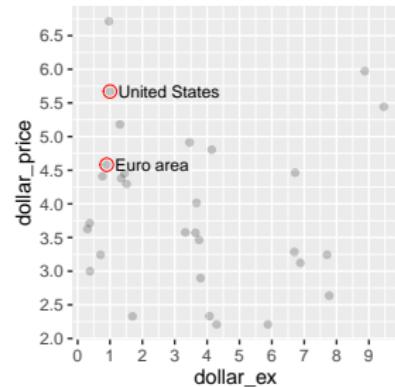
# Ejercicio

Vamos a enfatizar los puntos correspondientes a EEUU y Europa en el diagrama de dispersión de Bigmac.

- Crea una nueva tabla filtrando las filas correspondientes a EEUU y Europa

```
1 bigmac_usa_y_europa <- bigmac |>
2 filter(name %in% c("United States", "Euro area"))
```

- En el diagrama de dispersión, añade puntos en rojo para esa nueva tabla; usa un “geom\_point” adicional para bigmac\_usa\_y\_europa
- Haz que la fuente de datos de la capa “geom\_text\_repel” también sea de bigmac\_usa\_y\_europa



## 2 variables categóricas

- Cuando las variables tienen pocos niveles, usar “geom\_bar”
- Cuando las variables tienen muchos niveles, usar “geom\_tile”
- “geom\_bar” hace una transformación implícita (cuenta frecuencias), mientras que “geom\_tile” no lo hace

1. `geom_bar()` begins with the `diamonds` data set

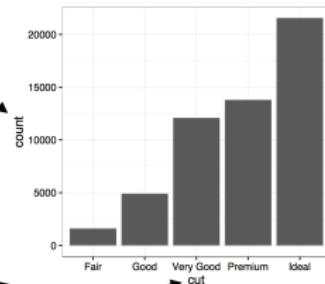
carat	cut	color	clarity	depth	table	price	x	y	z
0.23	Ideal	E	SI2	61.5	55	326	3.95	3.88	2.43
0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58	334	4.20	4.23	2.63
0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
...	...	...	...	...	...	...	...	...	...

2. `geom_bar()` transforms the data with the “count” stat, which returns a data set of cut values and counts.

`stat_count()`

cut	count	prop
Fair	1610	1
Good	4906	1
Very Good	12082	1
Premium	13791	1
Ideal	21551	1

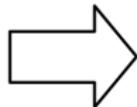
3. `geom_bar()` uses the transformed data to build the plot. cut is mapped to the x axis, count is mapped to the y axis.



## Fusionar y separar columnas

tidyverse tiene dos funciones para fusionar y separar columnas; se llaman “pivot\_longer” y “pivot\_wider”, respectivamente.

<b>id</b>	<b>bp1</b>	<b>bp2</b>
A	100	120
B	140	115
C	120	125



<b>id</b>	<b>measurement</b>	<b>value</b>
A	bp1	100
A	bp2	120
B	bp1	140
B	bp2	115
C	bp1	120
C	bp2	125

# Ejes, títulos y títulos de leyendas

- Usa “labs” para etiquetar ejes, títulos y títulos de leyendas
- Puedes romper textos largos en líneas (i) manualmente usando “\n” o (ii) automáticamente con “str\_wrap”
- Puedes eliminar todas las leyendas con “theme(legend.position = “none”)”

## Escalas de los ejes y las leyendas

- Las escalas se modifican con “scale”. Usa “break” para escalas numéricas y “labels” para categóricas
- El texto de las escalas se puede rotar con “theme(axis.text.x = element\_text(angle = 45, vjust = 0.5, hjust=0.5)”
- Las etiquetas de las variables categóricas pueden modificarse siempre con “fct\_recode” (incluídos los facets)
- La posición de las leyendas se cambia con “theme(legend.position = “left” )”
- Los colores se cambian con “scale\_color” . Por ejemplo, ver:  
<https://r4ds.hadley.nz/communication.html#fig-brewer>

# Layout

- Cambia el layout con “theme”.
- Temas predefinidos: <https://r4ds.hadley.nz/communication.html#fig-themes>
- Más temas: <https://jrnlold.github.io/ggthemes/>

# Más información sobre ggplot2

Los libros de Hadley Wickham:

- R for Data Science
- ggplot2: Elegant Graphics for Data Analysis

# Más información sobre cómo realizar buenas gráficas

- [Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures](#)
- [Graphical Data Analysis with R](#)
- [Truthful Art, The: Data, Charts, and Maps for Communication](#)