



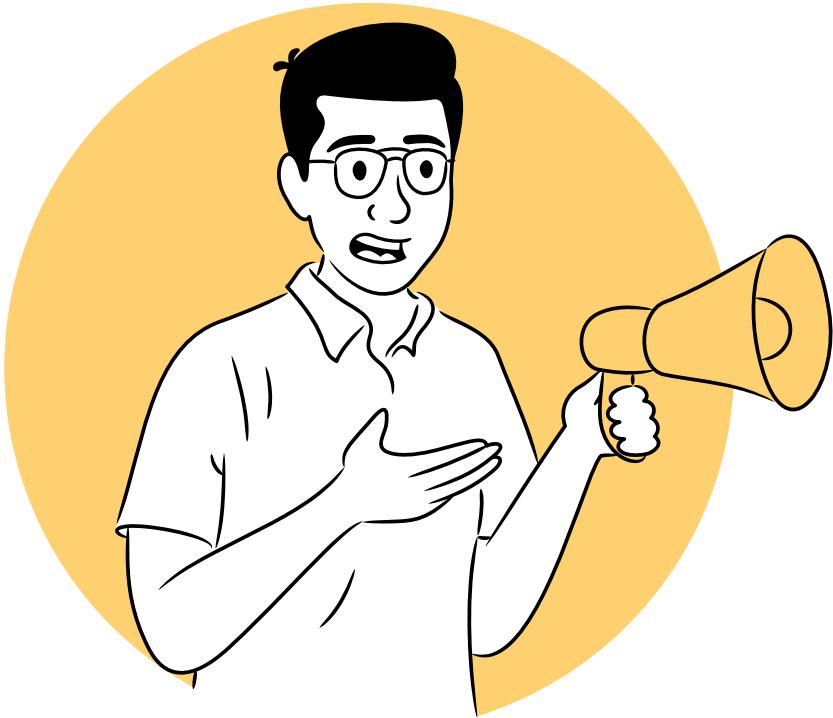
TALLER INTRODUCCIÓN A LA CIENCIA DE DATOS

Mtro. David Martínez Galicia

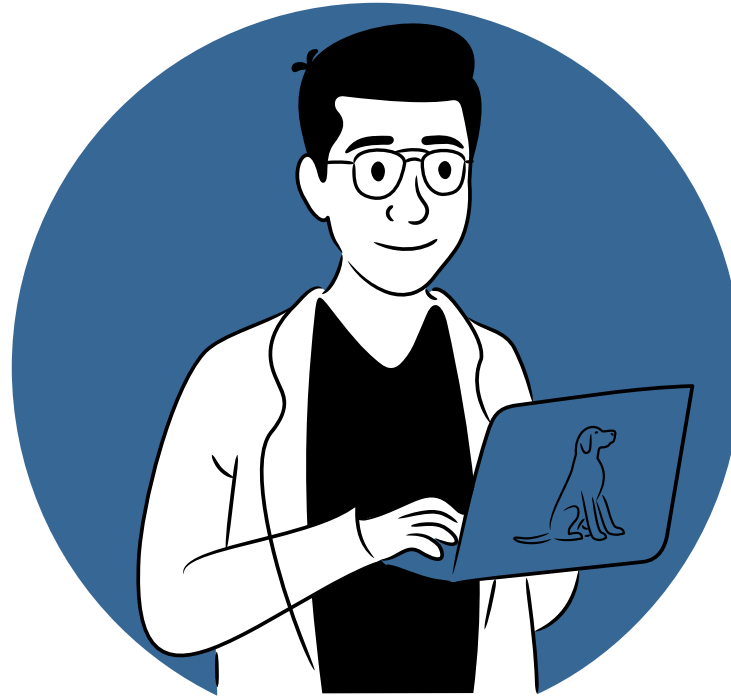
davidgalicia@outlook.es

Jornada Académica de Ingenierías
Universidad de Xalapa

Presentación



Divulgación



Docencia / Investigación



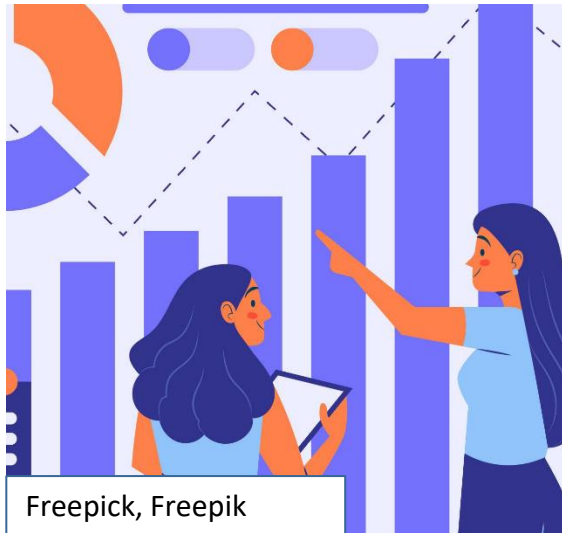
Creatividad

Agenda

- Ciencia de datos
- Dos pesitos de programación
- Entender el contexto
- Jugar con los datos
- Hacer las preguntas correctas
- Presentar los resultados

- La ciencia de datos es un **campo emergente** que se puede definir como la intersección de la computación, la estadística y diversos campos de aplicación.
- Tiene como objetivo extraer **información significativa**.

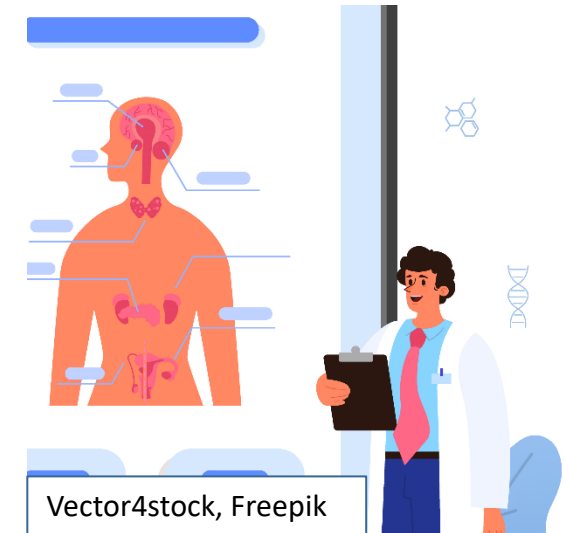
Ciencia de datos



Analítica de clientes
para entender y proveer
ofertas oportunas.



Detección de fraude
para identificar, rastrear y
prevenir el fraude.



Diagnóstico médico para
observar síntomas e
identificar patologías.

Dos pesitos de programación

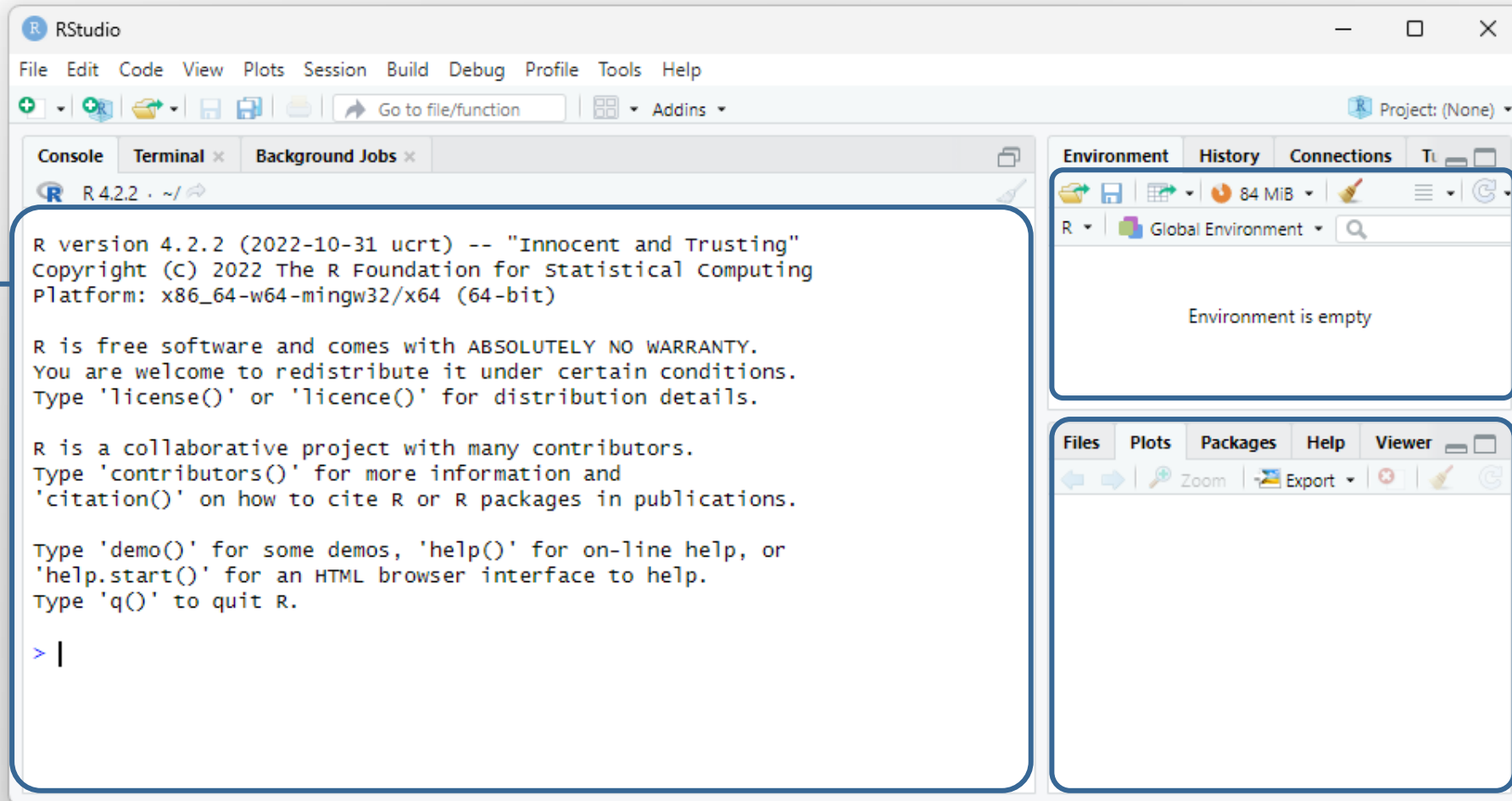
- R es un **lenguaje** de programación que nos permite **describir** los **pasos** para procesar los datos.
- Rstudio es un **entorno** de programación para R, es decir, es un conjunto de **herramientas** que nos facilita la programación.

Dos pesitos de programación

- Instrucciones, instaladores y datos.
- <https://goo.su/6SlekUI>

Dos pesitos de programación

1
Consola
ejecuta
código.



2
Ambiente
muestra los
datos.

3
Monitor
muestra las
gráficas.

Dos pesitos de programación



Nombre:
Juanita Hernández

Edad:
27 años

Donante de órganos:
Verdadero

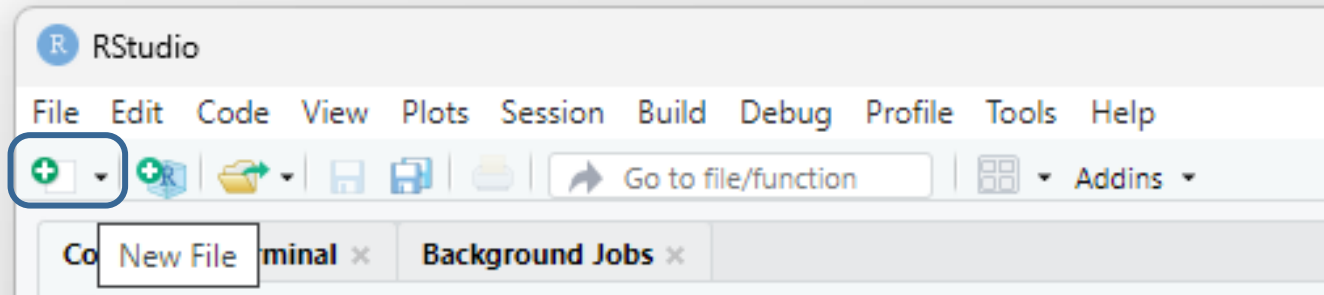
- Un dato es la **cantidad mínima** de información, por sí solo no tiene sentido.
- Existen varios **tipos** como: cadenas de texto, números, valores lógicos, entre otros.

Dos pesitos de programación

- **Variable:** espacio donde se almacena un dato.
- **Vector:** estructura en forma de lista que almacena datos.
- **Función:** código que realiza una tarea, puede recibir datos.
- **Librería:** Conjunto de funciones y herramientas con un objetivo.

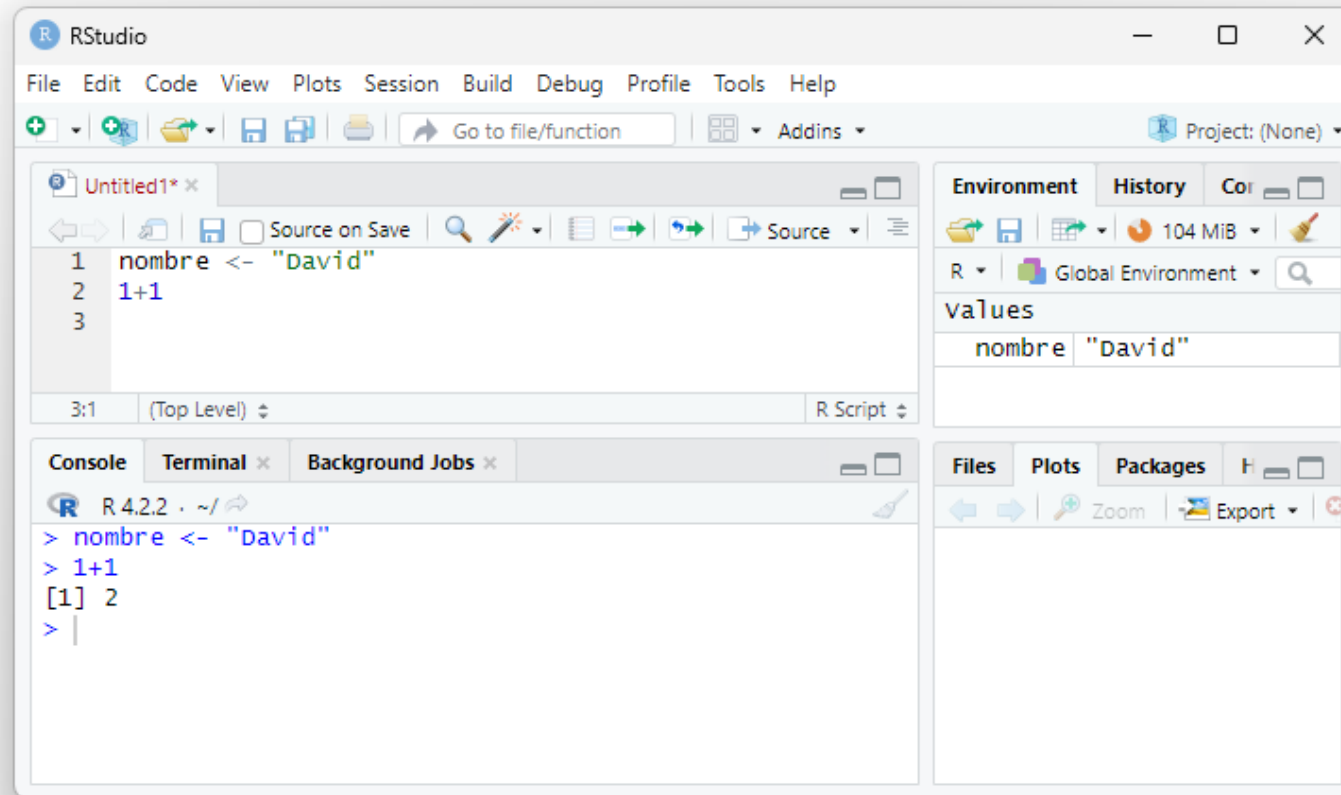
Dos pesitos de programación

- Paso 1: Crear un nuevo archivo o script.



Dos pesitos de programación

- **Ctrl + enter** = ejecutar línea.

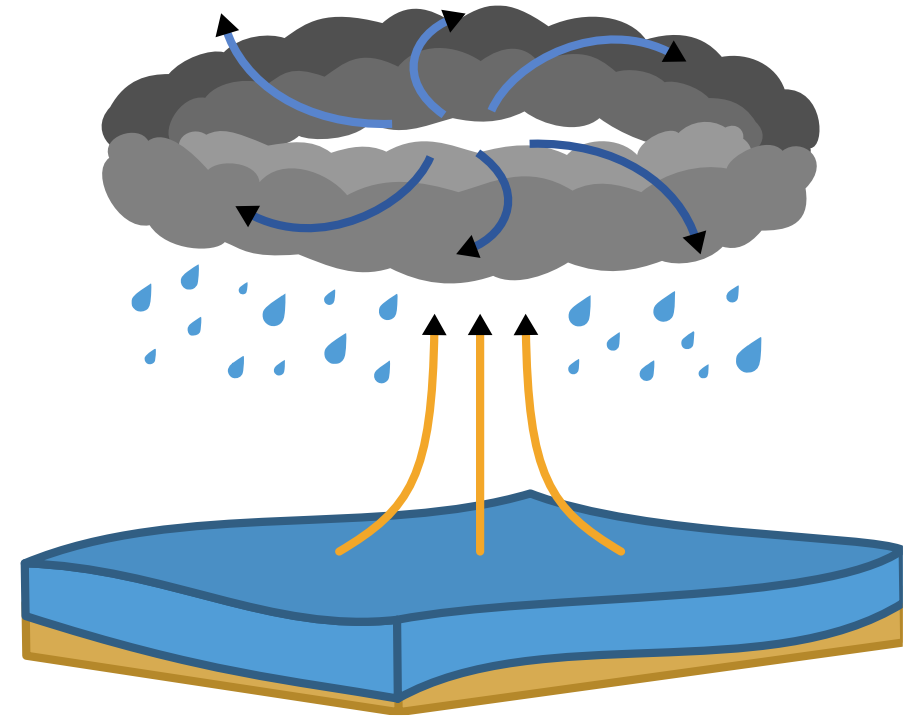


Entender el contexto

- Para poder analizar una base de datos hay que saber de **dónde provienen** los datos, su **significado** y tener mucha **curiosidad**.
- De ahora en adelante, seremos unos expertos en ciclones.

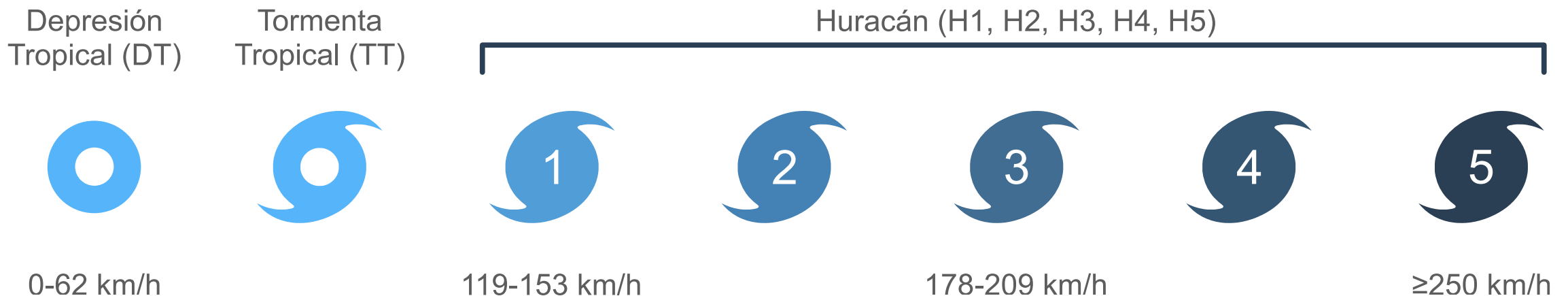
Entender el contexto

- Los ciclones son un **fenómeno** meteorológico que se producen cuando el **aire caliente y húmedo** en océanos tropicales se eleva para formar **tormentas eléctricas** que giran en **espiral**.



Entender el contexto

- A medida que un ciclón gira, atrae **más aire caliente** y se fortalece. Los ciclones se pueden nombrar dependiendo de la **velocidad** de sus vientos. En especial, los huracanes son los más destructivos.



Entender el contexto

- Aunque los ciclones pueden generar **grandes pérdidas** materiales y humanas, ayudan a **mitigar sequías** y **equilibrar la temperatura** global.
- La temporada comienza el 15 de mayo y termina el 30 de noviembre.



Entender el contexto

- La Oficina Nacional de Administración Oceánica y Atmosférica (NOAA) de los EE. UU. guarda **registro** de la **trayectoria** y **características** de los ciclones tropicales desde 1851.



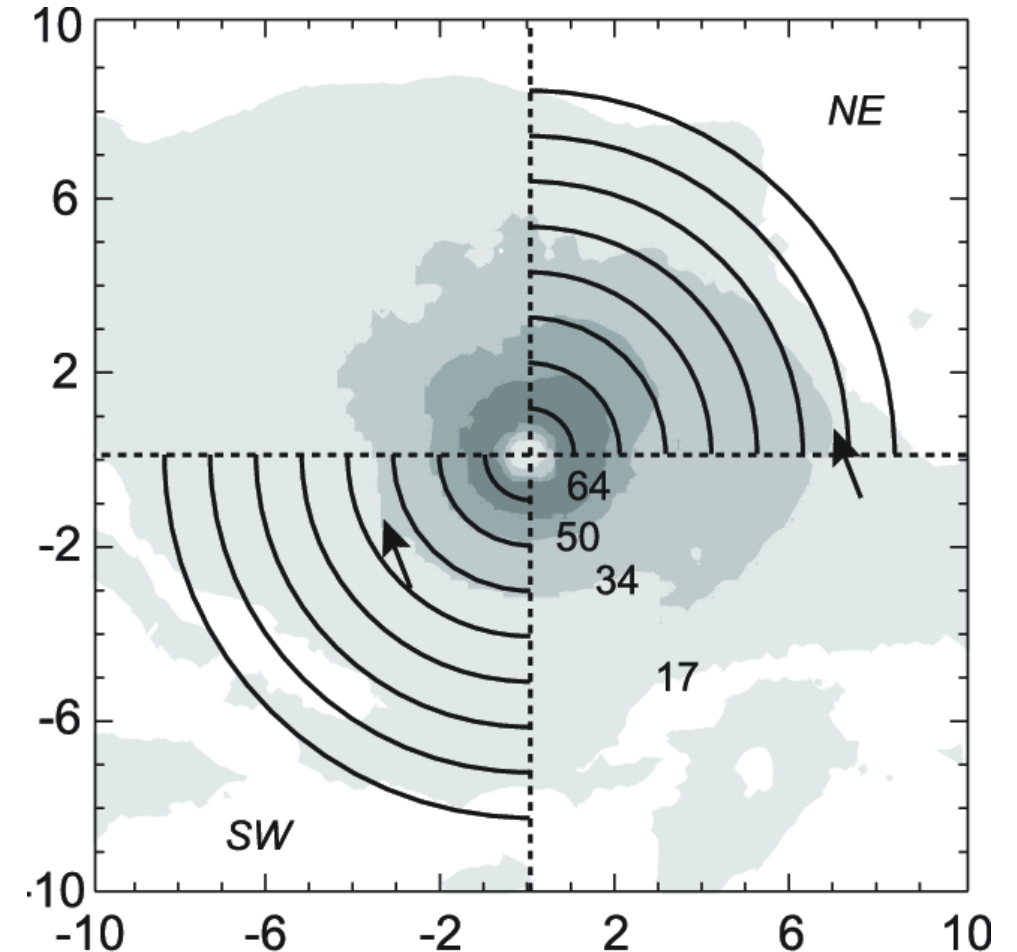
Entender el contexto

- La NOAA ofrece dos bases de datos de ciclones.
- Una para los ciclones del océano atlántico y otra para el Pacífico.
- <https://www.nhc.noaa.gov/data/#hurdat>
- En lo que respecta, ocuparemos una versión que contiene los datos de ambas bases.

Entender el contexto

¿Qué datos contiene?

- Nombre.
- Año, mes, día y hora.
- Latitud y longitud.
- Estado.
- Viento y presión.
- Radios de la tormenta.

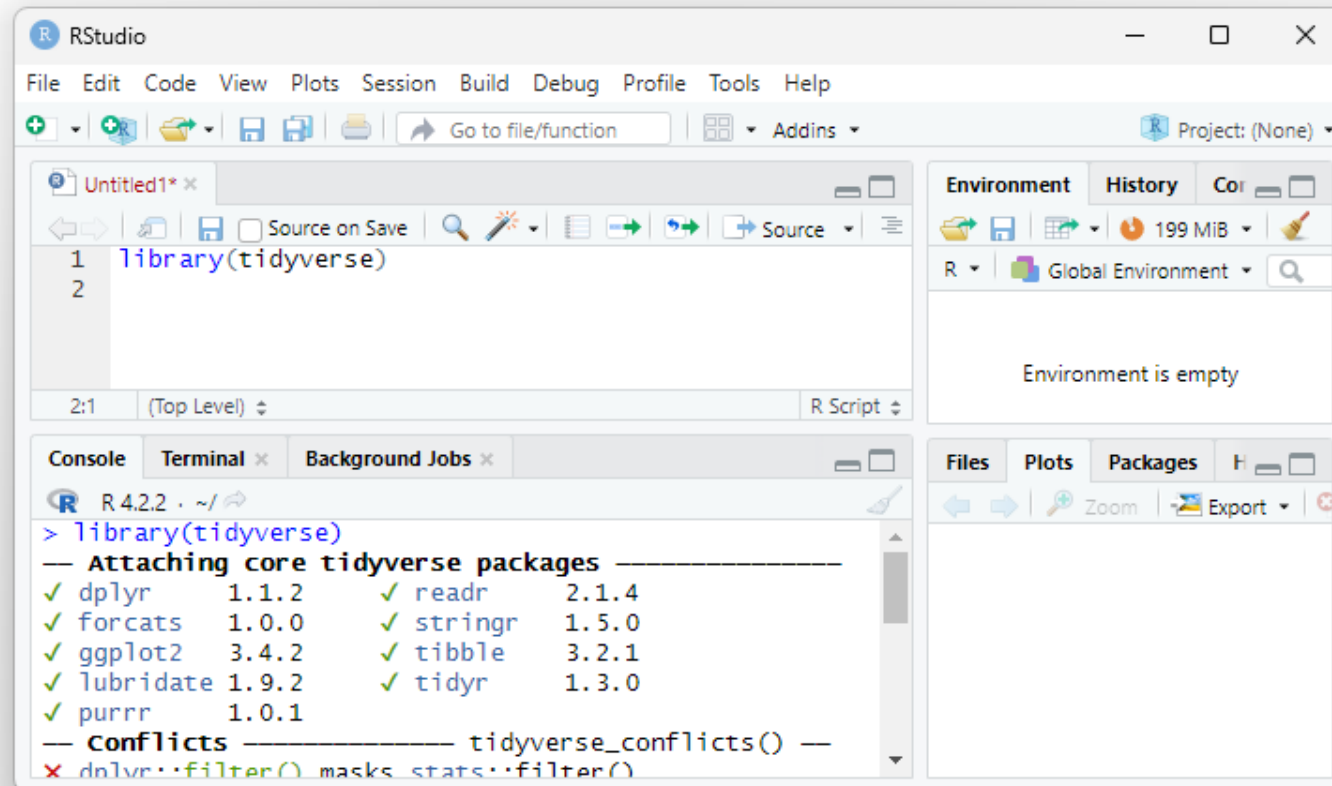


Jugar con los datos

- Para empezar a jugar necesitamos:
 1. Un archivo nuevo.
 2. Cargar las librerías que ocuparemos.
 3. Definir la carpeta en la que trabajaremos.
 4. Cargar los datos.

Jugar con los datos

- Crear un archivo nuevo, cargar **tidyverse** y ejecutar la línea.



The screenshot shows the RStudio interface with the following components:

- Source Editor:** Contains the code `library(tidyverse)` on line 1.
- Console:** Shows the output of the command, including a list of installed packages and a conflict warning.
- Environment:** Shows "Global Environment" and "Environment is empty".
- Files:** Shows the file explorer.
- Plots:** Shows the plot viewer.
- Packages:** Shows the installed packages.

Console Output:

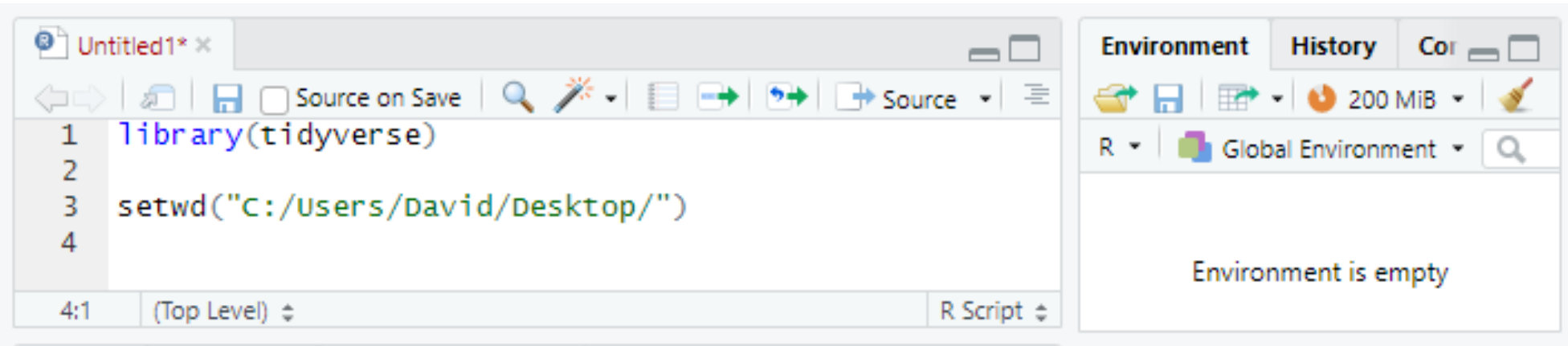
```
> library(tidyverse)
— Attaching core tidyverse packages —
✓ dplyr      1.1.2    ✓ readr      2.1.4
✓ forcats    1.0.0    ✓ stringr    1.5.0
✓ ggplot2     3.4.2    ✓ tibble     3.2.1
✓ lubridate  1.9.2    ✓ tidyr      1.3.0
✓ purrr       1.0.1
— Conflicts ————— tidyverse_conflicts() —
✖ dplyr::filter() masks stats::filter()
```

Jugar con los datos

- Crear una carpeta para guardar los datos y el archivo de R.
- Obtener su dirección, por ejemplo:
- C:\Users\David\Desktop\
- Definir la carpeta de trabajo con la función **setwd**.

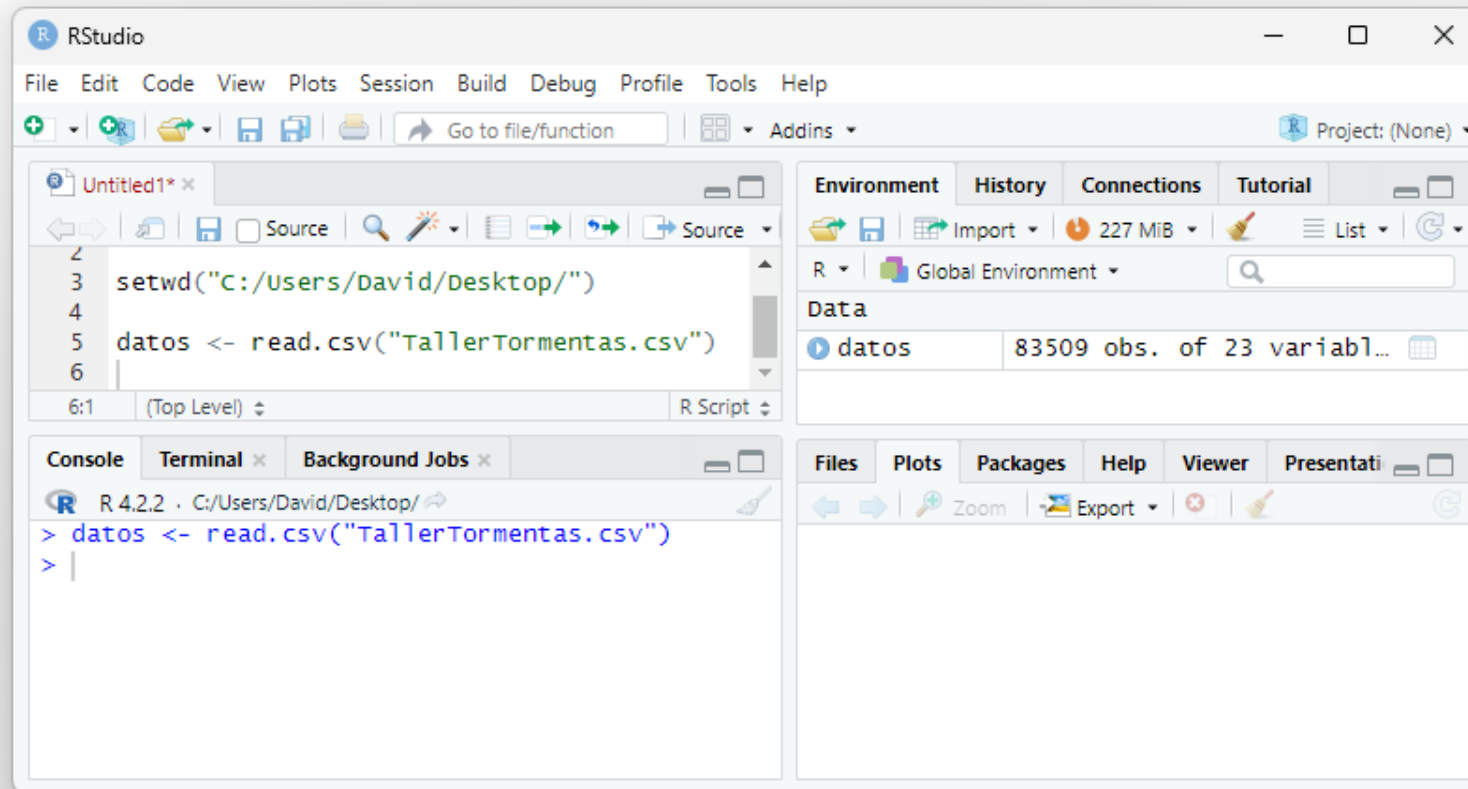
Jugar con los datos

- Definir la carpeta de trabajo con la función **setwd**.



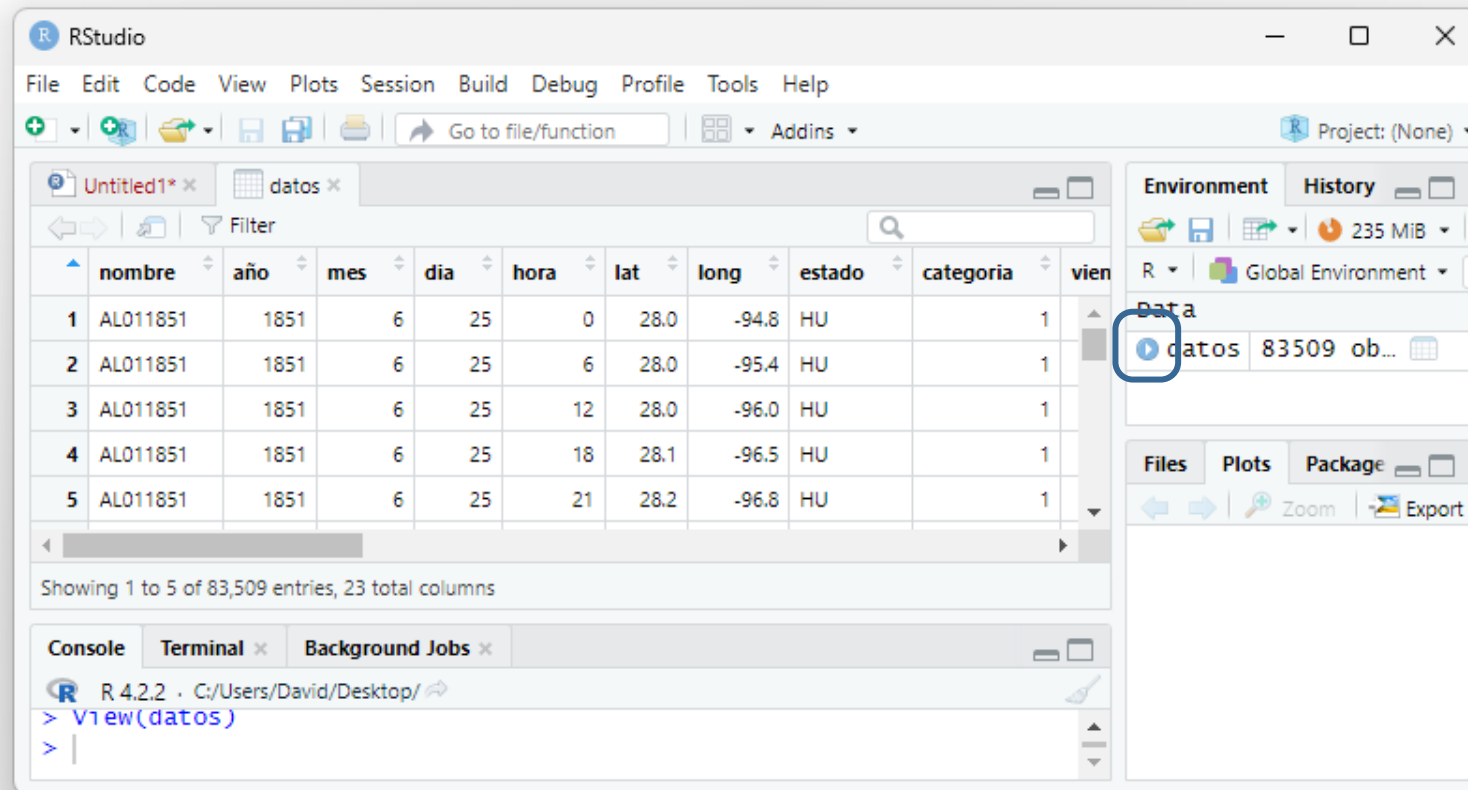
Jugar con los datos

- Leer y cargar datos usando la función **read.csv**.



Jugar con los datos

- Visualizar datos.



Jugar con los datos

- ¿Cuáles serían las primeras preguntas que te harías con estos datos?

Jugar con los datos

Te propongo tres:

- ¿Cuántos registros tenemos?
- ¿Cuántas tormentas han sido registradas?
- ¿Cuáles son los estados de los ciclones registrados?

Jugar con los datos

- Necesitaremos más funciones para contestar esta pregunta.
- El operador **\$** para seleccionar datos de una columna.
- El operador pype **%>%** para simplificar el código.
- La función **nrow** para contar el número de registros.
- La función **unique** para eliminar valores repetidos.
- La función **length** para saber el número de valores.

Jugar con los datos

¿Cuántos registros tenemos?

`nrow(datos)`

Resultado = 83,509

¿Cuántas tormentas han sido registradas?

```
datos$nombre %>%  
  unique() %>%  
  length()
```

Resultado = 1,843

¿Cuáles son los estados de los ciclones registrados?

```
datos$estado %>%  
  unique()
```

Resultado = HU, TS, EX, TD, LO, DB, SD, SS, WV, ET, PT, ST, TY

Jugar con los datos


- Necesitamos investigar más.
- ¿Acaso no existen nombres repetidos?
- ¿Qué significan los estados?

¿Acaso no existen nombres repetidos?

- Desde que inicia el año, los huracanes comienzan a ser nombrados en **orden alfabético**, alternando nombres masculinos y femeninos.
- Los nombres que empiezan con **q, u, x, y** ó **z** se excluyen de la lista por ser poco comunes.
- Las listas se **reutilizan** cada seis años, por eso es común que cada cierto tiempo los nombres se repitan.
- Si un huracán fue **muy devastador**, como Katrina o Patricia, los científicos retiran ese nombre y lo sustituyen por otro.

Jugar con los datos

¿Qué significan los estados?

- TD – Depresión tropical (< 34 nudos)
 - TS – Tormenta tropical (34-63 nudos)
 - HU – Huracán (> 64 nudos)
 - EX – Ciclón extratropical (de cualquier intensidad)
 - SD - Depresión subtropical (< 34 nudos)
 - SS - Tormenta subtropical (> 34 nudos)
 - LO: Sistema de baja de presión (de cualquier intensidad)
 - WV – Onda tropical (de cualquier intensidad)
 - DB – Perturbación (de cualquier intensidad)
- 

¿Es necesario enfocarse en todos los estados?

Jugar con los datos

- Visto lo visto, necesitamos modificar nuestra base de datos.
- ¿Cómo podemos evitar que se repitan los nombres?
- ¿Cómo podemos filtrar los registros por el estado?

¿Cómo podemos evitar que se repitan los nombres?

- Agregando al nombre de la tormenta el año.
- Se necesita la función **mutate** para modificar los datos.

```
datos <- datos %>%  
  mutate(nombre = paste(nombre, año, sep = ""))
```

Jugar con los datos

- Repetimos el proceso.

```
datos$nombre %>%  
  unique() %>%  
  length()
```

Resultado = 3,119

Jugar con los datos

¿Cómo podemos filtrar los registros por el estado?

- Se necesita la función **filter** para seleccionar los datos.

```
datos <- datos %>%  
  filter(estado == "TD" | estado == "TS" | estado == "HU")
```

Número de registros: 71,312

Jugar con los datos

- Hay ocasiones en las que necesitamos crear nuevas. En este caso, por ejemplo, nos hace falta la categoría de las tormentas.

```
datos <- datos %>%  
  mutate(categoria = cut(viento,  
    breaks = c(0,63,119,154,178,109,252,1000),  
    labels = c(-1, 0, 1, 2, 3, 4, 5),  
    include.lowest = TRUE, ordered = TRUE))
```

Jugar con los datos

- Para facilitar aún más las cosas, vamos a agregar a la variable estado el número de categoría, pero solo a los huracanes.

```
datos <- datos %>%  
  mutate(estado = if_else(estado == "HU",  
    paste("HU", categoria, sep = ""),  
    estado))
```


Hacer las preguntas correctas

Te propongo una última pregunta:

- ¿El número de tormentas y su intensidad han aumentado los últimos años?

Hacer las preguntas correctas

- Necesitaremos más funciones para contestar esta pregunta.
- La función **group_by** para agrupar registros .
- La función **summarize** modificar la base de datos.
- Primero determinaremos la categoría máxima de cada tormenta.

Hacer las preguntas correctas

```
datos2 <- datos %>%  
  group_by(año, nombre) %>%  
  summarise(maxCategoria =  
    ifelse(is.element("HU5", estado), "HU5",  
    ifelse(is.element("HU4", estado), "HU4",  
    ifelse(is.element("HU3", estado), "HU3",  
    ifelse(is.element("HU2", estado), "HU2",  
    ifelse(is.element("HU1", estado), "HU1",  
    ifelse(is.element("TS", estado), "TS",  
    ifelse(is.element("TD", estado), "TD", "NA"))))))) ,  
  .groups = "keep")
```

Hacer las preguntas correctas

```
datos3 <- datos2 %>%  
  group_by(año, maxCategoria) %>%  
  summarise(conteo = n(), .groups = "keep") %>%  
  filter(año >= 1850, año <= 2019) %>%  
  mutate(decada = floor(año/10)*10) %>%  
  group_by(decada, maxCategoria) %>%  
  summarise(conteo = sum(conteo), .groups = "keep")
```

Hacer las preguntas correctas

- Ordenar las categorías.

```
datos3 <- datos3 %>% mutate(maxCategoria =  
  factor(maxCategoria,  
    levels = c("NA", "TD", "TS", "HU1", "HU2", "HU3", "HU4", "HU5")))
```

Presentar resultados

Hay dos opciones:

- Usar tablas.
- Usar gráficos.

	decada	maxCategoria	conteo
1	1850	HU4	9
2	1850	HU3	12
3	1850	HU2	18
4	1850	TS	14
5	1860	HU4	4
6	1860	HU3	28
7	1860	HU2	17
8	1860	TS	23
9	1870	HU4	14
10	1870	HU3	14
11	1870	HU2	27
12	1870	TS	20
13	1880	HU4	15
14	1880	HU3	18
15	1880	HU2	30
16	1880	TS	26
17	1890	HU4	16

Showing 1 to 17 of 83 entries, 3 total columns

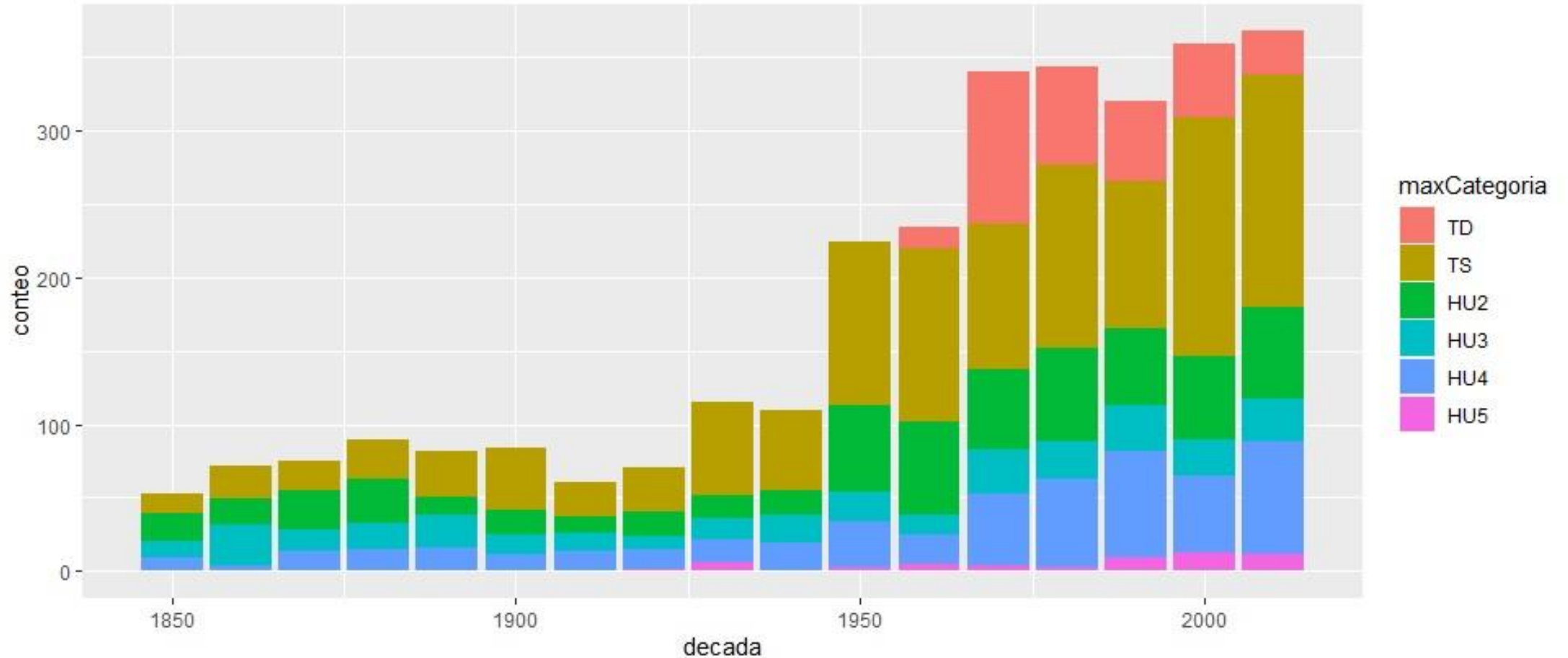
Presentar resultados

- Usaremos una gráfica de barras.

```
datos3 %>%
```

```
  ggplot(aes(x = decada, y = conteo, fill = maxCategoria)) +  
  geom_col()
```

Presentar resultados

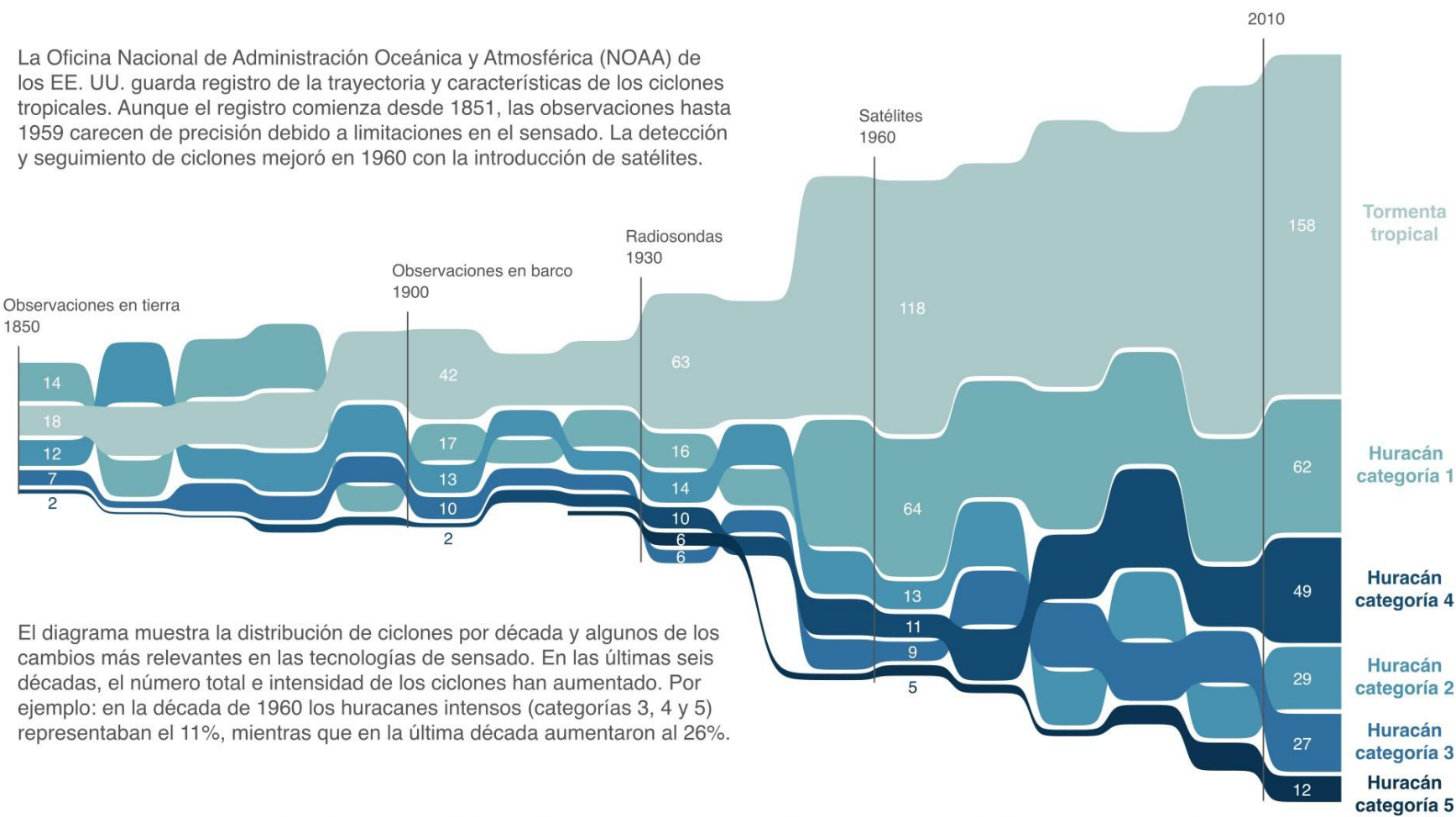


¿Qué crees que sugiere la gráfica?

Presentar resultados

CICLONES A TRAVÉS DE LAS DÉCADAS

La Oficina Nacional de Administración Oceánica y Atmosférica (NOAA) de los EE. UU. guarda registro de la trayectoria y características de los ciclones tropicales. Aunque el registro comienza desde 1851, las observaciones hasta 1959 carecen de precisión debido a limitaciones en el sensado. La detección y seguimiento de ciclones mejoró en 1960 con la introducción de satélites.



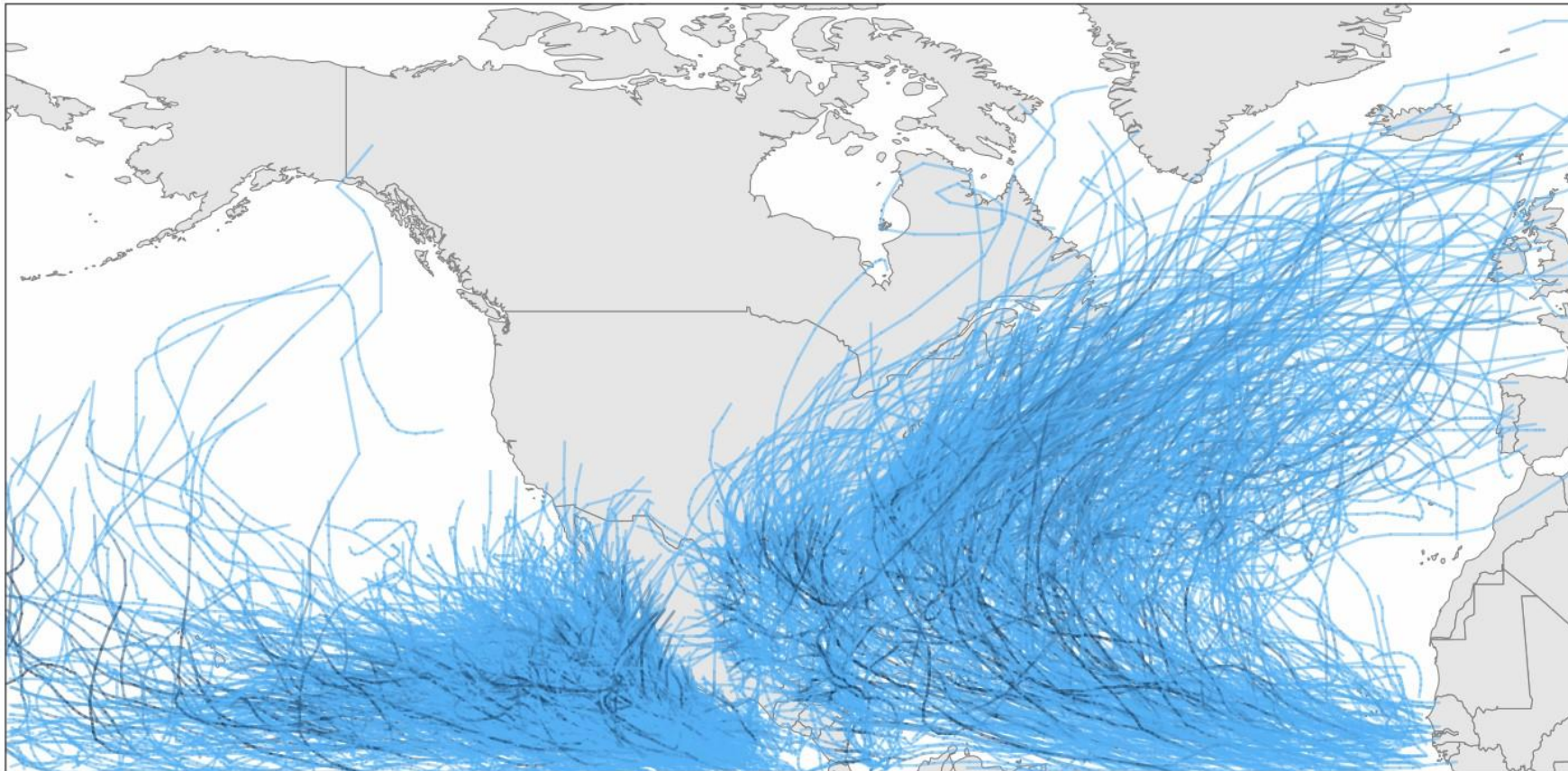
El diagrama muestra la distribución de ciclones por década y algunos de los cambios más relevantes en las tecnologías de sensado. En las últimas seis décadas, el número total e intensidad de los ciclones han aumentado. Por ejemplo: en la década de 1960 los huracanes intensos (categorías 3, 4 y 5) representaban el 11%, mientras que en la última década aumentaron al 26%.

Autor: David Martínez-Galicia | Twitter: @OyeDavidGalicia | Datos: Atlantic and NE/NC Pacific HURDAT2 (NHC, NOAA) | Paleta: Hokusai2 @ MetBrewer

Presentar resultados

Trayectoria de ciclones desde 1956 / Cyclone tracks since 1956

Categoría / Category DT-TT/TD-TS H1 H2 H3 H4 H5



Autor / Author: David Martínez-Galicia | Twitter: @OyeDavidGalicia | Datos / Data: Atlantic and NE/NC Pacific HURDAT2 (NHC, NOAA)

Presentar resultados

¿CUÁLES SON LOS ESTADOS MÁS AFECTADOS POR CICLONES?

Doscientos cincuenta ciclones azotaron México desde 1970 hasta 2021. El número de estados afectados por cada ciclón varía según su fuerza y trayectoria. El mapa divide los ciclones en dos categorías y muestra su número por estado.

Ingreso: Representa a ciclones cuyos centros golpean el estado después de dejar el océano.

Otros: Representa a ciclones cuyos centros no golpearon el estado o a ciclones que afectaron otros estados después de su ingreso en tierra.

¿CÓMO INTERPRETAR EL MAPA?



Autor: David Martínez-Galicia | Twitter: @OyeDavidGalicia
Datos: Ciclones que han impactado en México (SEMARNAT) | Paleta: Hokusai2 @ MetBrewer

