# # Conceptos básicos de ggplot

**{#ggplot-basics}**

**ggplot2** es el paquete de R más popular para la visualización de datos. Su función ggplot() es el núcleo de este paquete, y todo este enfoque se conoce coloquialmente como "ggplot", con las figuras resultantes a veces llamadas cariñosamente "ggplots". El "gg" en estos nombres refleja la "gramática de los gráficos" utilizada para construir las figuras. **ggplot2** se beneficia de una amplia variedad de paquetes R complementarios que mejoran aún más su funcionalidad.

La sintaxis es significativamente diferente de **la de** R, y tiene una curva de aprendizaje asociada. El uso de **ggplot2** generalmente requiere que el usuario formatee sus datos de una manera que sea altamente compatible con **tidyverse**, lo que en última instancia hace que el uso de estos paquetes juntos sea muy eficaz.

En esta página cubriremos los fundamentos del trazado con **ggplot2**. Vea la página de [consejos de ggplot](#ggplot-tips) para sugerencias y técnicas avanzadas para hacer que sus gráficos se vean realmente bien.

Hay varios tutoriales extensos de **ggplot2** enlazados en la sección de recursos. También puede descargar esta [hoja de trucos de visualización de datos con ggplot](https://github.com/rstudio/cheatsheets/raw/master/data-visualization-2.1.pdf) desde el sitio web de RStudio. Si quieres inspirarte en formas de visualizar tus datos de forma creativa, te sugerimos que revises sitios web como la [galería de gráficos de R](https://www.r-graph-gallery.com/) y [Data-to-viz](https://www.data-to-viz.com/caveats.html).

## Preparación

### Cargar paquetes

Este trozo de código muestra la carga de los paquetes necesarios para los análisis. En este manual destacamos p\_load() de **pacman**, que instala el paquete si es necesario y lo carga para su uso. También puede cargar los paquetes instalados con library() de **.** Consulta la página sobre [los fundamentos de R](#r-basics) para obtener más información sobre los paquetes de R.

### Importar datos

Importamos el conjunto de datos de casos de una epidemia de ébola simulada. Si quiere seguir el proceso, [clica para descargar linelist "limpia"](https://github.com/epirhandbook/Epi_R_handbook/raw/master/data/case_linelists/linelist_cleaned.rds) (como archivo .rds). Importe sus datos con la función import() del paquete **rio** (acepta muchos tipos de archivos como .xlsx, .rds, .csv - vea la página de [importación y exportación](#import-and-export) para más detalles).

A continuación se muestran las primeras 50 filas del listado. Nos centraremos en las variables continuas edad, wt\_kg (peso en kilos), ct\_blood (valores de CT) y days\_onset\_hosp (diferencia entre la fecha de inicio y la hospitalización).

### Limpieza general

Cuando se preparan los datos para trazarlos, lo mejor es hacer que los datos se adhieran a [los estándares de datos "ordenados" tanto como](https://r4ds.had.co.nz/tidy-data.html) sea posible. En las páginas de gestión de datos de este manual, como [Limpieza de datos y funciones básicas, se explica](#cleaning-data-and-core-functions) cómo conseguirlo.

Algunas formas sencillas de preparar nuestros datos para que sean mejores para el trazado pueden incluir la mejora del contenido de los datos para su visualización, lo que no equivale necesariamente a una mejor manipulación de los datos. Por ejemplo:

* Sustituir los valores NA de una columna de caracteres por la cadena de caracteres "Desconocido"
* Considere la posibilidad de convertir la columna en de tipo factor para que sus valores tengan niveles ordinales prescritos
* Limpiar algunas columnas para que sus valores "amigables con los datos" con barra baja, etc., se cambien a texto normal o a mayúsculas y minúsculas (ver [Caracteres y cadenas](#characters-and-strings))

He aquí algunos ejemplos de esto en acción:

### Pivotar más tiempo

Como una cuestión de estructura de datos, para **ggplot2** a menudo también queremos pivotar nuestros datos en formatos más largos. Lea más sobre esto en la página de [Pivoteo de datos](#pivoting-data).

Por ejemplo, digamos que queremos trazar datos que están en un formato "amplio", como por ejemplo para cada caso en linelist y sus síntomas. A continuación creamos una minilista llamada datos\_síntomas que contiene sólo las columnas case\_id y symptoms.

Así es como se ven las primeras 50 filas de esta mini-lista - vea cómo están formateadas "a lo ancho" con cada síntoma como una columna:

Si quisiéramos trazar el número de casos con síntomas específicos, estamos limitados por el hecho de que cada síntoma es una columna específica. Sin embargo, podemos hacer pivotar las columnas de síntomas a un formato más largo como este:

Aquí están las primeras 50 filas. Observe que el caso tiene 5 filas - una para cada síntoma posible. Las nuevas columnas symptom\_name y symptom\_is\_present son el resultado del pivote. Ten en cuenta que este formato puede no ser muy útil para otras operaciones, pero es útil para trazar.

## Fundamentos de ggplot

**"Gramática de los gráficos" - ggplot2**

El trazado con **ggplot2** se basa en "añadir" capas de trazado y elementos de diseño unos sobre otros, añadiendo cada comando a los anteriores con un símbolo de suma (+). El resultado es un objeto de trazado multicapa que se puede guardar, modificar, imprimir, exportar, etc.

Los objetos ggplot pueden ser muy complejos, pero el orden básico de las capas suele ser el siguiente:

1. Comience con el comando ggplot() de la línea de base - esto "abre" el ggplot y permite que las funciones subsecuentes sean agregadas con +. Normalmente, el conjunto de datos también se especifica en este comando
2. Añadir capas "geom" - estas funciones visualizan los datos como geometrías (formas), por ejemplo, como un gráfico de barras, un gráfico de líneas, un gráfico de dispersión, un histograma (¡o una combinación!). Todas estas funciones comienzan con geom\_ como prefijo.
3. Añada elementos de diseño al gráfico, como etiquetas de ejes, título, fuentes, tamaños, esquemas de color, leyendas o rotación de ejes.

Un ejemplo sencillo de código esqueleto es el siguiente. Explicaremos cada componente en las secciones siguientes.

## ggplot()

El comando de apertura de cualquier gráfico ggplot2 es ggplot(). Este comando simplemente crea un lienzo en blanco sobre el que añadir capas. Se "abre" el camino para añadir más capas con un símbolo +.

Normalmente, el comando ggplot() incluye el argumento data = para el gráfico. Esto establece el conjunto de datos por defecto que se utilizará para las capas posteriores del gráfico.

Este comando terminará con un + después de su paréntesis de cierre. Esto deja el comando "abierto". El ggplot sólo se ejecutará/aparecerá cuando el comando completo incluya una capa final sin un + al final.

## Geoms

Un lienzo en blanco no es suficiente: necesitamos crear geometrías (formas) a partir de nuestros datos (por ejemplo, gráficos de barras, histogramas, gráficos de dispersión, gráficos de caja).

Esto se hace añadiendo capas "geoms" al comando inicial ggplot(). Hay muchas funciones de **ggplot2** que crean "geoms". Cada una de estas funciones comienza con "geom\_", por lo que nos referiremos a ellas genéricamente como geom\_XXXX(). Hay más de 40 geoms en **ggplot2** y muchos otros creados por aficionados. Véalos en la [galería de ggplot2](https://exts.ggplot2.tidyverse.org/gallery/). Algunos geoms comunes se enumeran a continuación:

* Histogramas - geom\_histograma()
* Gráficos de barras - geom\_bar() o geom\_col() (véase [la sección "Gráfico de barras"](#ggplot_basics_bars))
* Gráficos de caja - geom\_boxplot()
* Puntos (por ejemplo, gráficos de dispersión) - geom\_point()
* Gráficos de líneas - geom\_line() o geom\_path()
* Líneas de tendencia - geom\_smooth()

En un gráfico se pueden mostrar uno o varios geoms. Cada uno se añade a los comandos anteriores **de ggplot2** con un +, y se trazan secuencialmente de manera que los geoms posteriores se trazan encima de los anteriores.

## Asignación de datos al gráfico

A la mayoría de las funciones geom hay que decirles qué utilizar para crear sus formas, por lo que hay que indicarles cómo deben asignar las columnas de los datos a los componentes del gráfico, como los ejes, los colores de las formas, los tamaños de las formas, etc. Para la mayoría de las funciones geom, los componentes esenciales que deben asignarse a las columnas de los datos son el eje x y (si es necesario) el eje y.

Este "mapeo" se produce con el argumento mapping =. Los mapeos que proporcione a mapping deben estar envueltos en la función aes(), por lo que escribiría algo como mapping = aes(x = col1, y = col2), como se muestra a continuación.

A continuación, en el comando ggplot() los datos se establecen como linelist de casos. En el argumento mapping = aes() la columna edad se asigna al eje x, y la columna wt\_kg se asigna al eje y.

Después de un +, los comandos de trazado continúan. Se crea una forma con la función "geom" geom\_point(). Este geom hereda los mapeos del comando ggplot() anterior - conoce las asignaciones eje-columna y procede a visualizar esas relaciones como puntos en el lienzo.

Como otro ejemplo, los siguientes comandos utilizan los mismos datos, un mapeo ligeramente diferente y un geom diferente. La función geom\_histograma() sólo requiere una columna mapeada en el eje x, ya que el eje y de las cuentas se genera automáticamente.

### Estética de la trama

En la terminología de ggplot, la "estética" de un gráfico tiene un significado específico. Se refiere a una propiedad visual de los datos trazados. Ten en cuenta que "estética" aquí se refiere a los datos que se trazan en geoms / formas - no la pantalla circundante, tales como títulos, etiquetas de los ejes, el color de fondo, que se podría asociar con la palabra "estética" en Inglés común. En ggplot esos detalles se llaman "temas" y se ajustan dentro de un comando theme() (ver [esta sección](#ggplot_basics_themes)).

Por lo tanto, la estética del objeto de ploteo puede ser colores, tamaños, transparencias, colocación, etc. de los datos ploteados. No todos los geoms tendrán las mismas opciones estéticas, pero muchas pueden ser utilizadas por la mayoría de los geoms. He aquí algunos ejemplos:

* shape = Mostrar un punto con geom\_point() como punto, estrella, triángulo o cuadrado...
* fill = El color interior (por ejemplo, de una barra o boxplot)
* color = La línea exterior de una barra, boxplot, etc., o el color del punto si se utiliza geom\_point()
* size = Tamaño (por ejemplo, grosor de línea, tamaño de punto)
* alpha = Transparencia (1 = opaco, 0 = invisible)
* binwidth = Ancho de los bins del histograma
* anchura = Anchura de las columnas del "diagrama de barras"
* linetype = Tipo de línea (por ejemplo, sólida, discontinua, punteada)

A esta estética de los objetos de la trama se le pueden asignar valores de dos maneras:

1. Se asigna un valor estático (por ejemplo, color = "azul") que se aplica a todas las observaciones trazadas
2. Se asigna a una columna de los datos (por ejemplo, color = hospital) de manera que la visualización de cada observación depende de su valor en esa columna

### Establecer un valor estático

Si se desea que la estética del objeto de trazado sea estática, es decir, que sea la misma para cada observación de los datos, se escribe su asignación dentro del geom pero fuera de cualquier sentencia mapping = aes(). Estas asignaciones podrían ser como size = 1 o color = "blue". Aquí hay dos ejemplos:

* En el primer ejemplo, el mapeo = aes() está en el comando ggplot() y los ejes se asignan a las columnas de edad y peso en los datos. La estética del gráfico color =, tamaño =, y alfa = (transparencia) se asignan a valores estáticos. Para mayor claridad, esto se hace en la función geom\_point(), ya que se pueden añadir otros geoms después que tomarían valores diferentes para su estética de trazado.
* En el segundo ejemplo, el histograma requiere sólo el eje x mapeado a una columna. La anchura del histograma =, el color =, el relleno = (color interno), y el alfa = se establecen de nuevo dentro del geom a valores estáticos.

### Escalado a los valores de la columna

La alternativa es escalar la estética del objeto de trazado por los valores de una columna. En este enfoque, la visualización de esta estética dependerá del valor de esa observación en esa columna de los datos. Si los valores de la columna son continuos, la escala de visualización (leyenda) para esa estética será continua. Si los valores de la columna son discretos, la leyenda mostrará cada valor y los datos trazados aparecerán como claramente "agrupados" (lea más en la sección de [agrupación](#ggplotgroups) de esta página).

Para conseguirlo, se asigna esa estética de gráfico a un nombre de columna (sin comillas). Esto debe hacerse dentro de una función *mapping = aes()* (nota: hay varios lugares en el código donde puedes hacer estas asignaciones de mapeo, como se discute [a continuación](##ggplot_basics_map_loc)).

A continuación, dos ejemplos.

* En el primer ejemplo, el color = estético (de cada punto) está mapeado a la columna edad - ¡y ha aparecido una escala en una leyenda! Por ahora sólo hay que tener en cuenta que la escala existe - mostraremos cómo modificarla en secciones posteriores.
* En el segundo ejemplo, dos nuevas estéticas de trazado también se asignan a columnas (color = y tamaño =), mientras que las estéticas de trazado forma = y alfa = se asignan a valores estáticos fuera de cualquier función de mapeo = aes().

Nota: Las asignaciones de los ejes siempre se asignan a las columnas de los datos (no a los valores estáticos), y esto se hace siempre dentro de mapping = aes().

Es importante mantener un seguimiento de las capas y la estética de los gráficos cuando se hacen gráficos más complejos, por ejemplo, gráficos con múltiples geom. En el ejemplo siguiente, el tamaño = estético se asigna dos veces - una para geom\_point() y otra para geom\_smooth() - ambas veces como un valor estático.

### Dónde hacer las asignaciones cartográficas

El mapeo estético dentro de mapping = aes() puede escribirse en varios lugares en sus comandos de trazado e incluso puede escribirse más de una vez. Esto puede ser escrito en el comando ggplot() superior, y/o para cada geom individual debajo. Los matices incluyen:

* Las asignaciones de mapeo realizadas en el comando ggplot() superior se heredarán por defecto en cualquier geom a continuación, al igual que se heredan x = e y =
* Las asignaciones cartográficas realizadas dentro de un geom se aplican sólo a ese geom

Del mismo modo, los datos = especificados en la parte superior de ggplot() se aplicarán por defecto a cualquier geom de abajo, pero también se podrían especificar datos para cada geom (pero esto es más difícil).

Así, cada uno de los siguientes comandos creará el mismo gráfico:

### Grupos

Puedes agrupar fácilmente los datos y "trazar por grupo". De hecho, ¡ya lo has hecho!

Asignar la columna "agrupación" a la estética de la gráfica adecuada, dentro de un mapeo = aes(). Arriba, hemos demostrado esto usando valores continuos cuando asignamos el tamaño del punto = a la columna edad. Sin embargo, esto funciona de la misma manera para las columnas discretas/categóricas.

Por ejemplo, si quiere que los puntos se muestren por género, deberá establecer mapeo = aes(color = género). Automáticamente aparecerá una leyenda. Esta asignación puede hacerse dentro de mapping = aes() en el comando ggplot() superior (y ser heredado por el geom), o podría establecerse en un mapping = aes() separado dentro del geom. Ambos enfoques se muestran a continuación:

Ten en cuenta que dependiendo de la geom, tendrá que utilizar diferentes argumentos para agrupar los datos. Para geom\_point() lo más probable es que utiliza color =, forma = o tamaño =. Mientras que para geom\_bar() es más probable que utiliza fill =. Esto sólo depende del geom y de la estética del gráfico que desee para reflejar las agrupaciones.

Para su información - la forma más básica de agrupar los datos es utilizando sólo el argumento group = dentro de mapping = aes(). Sin embargo, esto por sí mismo no cambiará los colores, el relleno o las formas. Tampoco creará una leyenda. Sin embargo, los datos están agrupados, por lo que las visualizaciones estadísticas pueden verse afectadas.

Para ajustar el orden de los grupos en un gráfico, consulta la página de [consejos de ggplot](#ggplot-tips) o la página sobre [Factores](#factors). Hay muchos ejemplos de gráficos agrupados en las secciones siguientes sobre el trazado de datos continuos y categóricos.

## Facetas / Múltiplos pequeños

Las facetas, o "pequeños múltiplos", se utilizan para dividir un gráfico en una figura de varios paneles, con un panel ("faceta") por grupo de datos. El mismo tipo de gráfico se crea varias veces, cada una de ellas utilizando un subgrupo del mismo conjunto de datos.

El facetado es una funcionalidad que viene con **ggplot2**, por lo que las leyendas y los ejes de los "paneles" facetados se alinean automáticamente. Hay otros paquetes discutidos en la página de [consejos de ggplot](#ggplot-tips) que se utilizan para combinar gráficos completamente diferentes (**cowplot** y **patchwork**) en una figura.

El facetado se realiza con una de las siguientes funciones de **ggplot2**:

1. facet\_wrap() Para mostrar un panel diferente para cada nivel de una sola variable. Un ejemplo de esto podría ser mostrar una curva de epidemia diferente para cada hospital de una región. Las facetas se ordenan alfabéticamente, a menos que la variable sea un factor con otro orden definido.

* Puedes invocar ciertas opciones para determinar la disposición de las facetas, por ejemplo, nrow = 1 o ncol = 1 para controlar el número de filas o columnas en las que se organizan los gráficos con facetas.

1. facet\_grid() Se utiliza cuando se quiere introducir una segunda variable en la disposición de las facetas. Aquí cada panel de una cuadrícula muestra la intersección entre los valores de dos columnas. Por ejemplo, las curvas epidémicas para cada combinación hospital-grupo de edad con los hospitales en la parte superior (columnas) y los grupos de edad en los lados (filas).

* nrow y ncol no son relevantes, ya que los subgrupos se presentan en una cuadrícula

Cada una de estas funciones acepta una sintaxis de fórmula para especificar la(s) columna(s) para el facetado. Ambas aceptan hasta dos columnas, una a cada lado de la tilde ~.

* Para facet\_wrap() lo más frecuente es escribir una sola columna precedida de una tilde ~ como facet\_wrap(~hospital). Sin embargo, puede escribir dos columnas facet\_wrap(resultado ~hospital) - cada combinación única se mostrará en un panel separado, pero no se organizarán en una cuadrícula. Los encabezados mostrarán los términos combinados y éstos no tendrán una lógica específica para las columnas frente a las filas. Si sólo proporciona una variable de facetado, se utiliza un punto . como marcador de posición en el otro lado de la fórmula - vea los ejemplos de código.
* Para facet\_grid() también puede especificar una o dos columnas a la fórmula (filas de la cuadrícula ~ columnas). Si sólo quiere especificar una, puede colocar un punto . al otro lado de la tilde como facet\_grid(. ~ hospital) o facet\_grid(hospital ~ .).

Las facetas pueden contener rápidamente una cantidad abrumadora de información, por lo que conviene asegurarse de no tener demasiados niveles de cada variable por la que se elija la faceta. He aquí algunos ejemplos rápidos con el conjunto de datos sobre la malaria (véase [Descargar el manual y los datos](#download-handbook-and-data)), que consiste en el recuento diario de casos de malaria en los centros, por grupos de edad.

A continuación importamos y hacemos algunas modificaciones rápidas para simplificar:

A continuación se muestran las primeras 50 filas de los datos sobre la malaria. Obsérvese que hay una columna malaria\_tot, pero también columnas para los recuentos por grupo de edad (que se utilizarán en el segundo ejemplo de facet\_grid()).

### facet\_wrap()

Por el momento, vamos a centrarnos en las columnas malaria\_tot y Distrito. Ignoremos por ahora las columnas de recuento por edad. Trazaremos las curvas epidémicas con geom\_col(), que produce una columna para cada día a la altura del eje y especificada en la columna malaria\_tot (los datos ya son recuentos diarios, por lo que utilizamos geom\_col() - véase [la sección "Diagrama de barras" más adelante](#ggplot_basics_bars)).

Cuando añadimos el comando facet\_wrap(), especificamos una tilde y a continuación la columna sobre la que hacer la faceta (Distrito en este caso). Puedes colocar otra columna a la izquierda de la tilde, - esto creará una faceta para cada combinación - pero le recomendamos que lo haga con facet\_grid() en su lugar. En este caso de uso, se crea una faceta para cada valor único de Distrito.

### facet\_grid()

Podemos utilizar un enfoque de facet\_grid() para cruzar dos variables. Digamos que queremos cruzar Distrito y edad. Bien, necesitamos hacer algunas transformaciones de datos en las columnas de edad para poner estos datos en el formato "largo" preferido por ggplot. Los grupos de edad tienen sus propias columnas - los queremos en una sola columna llamada age\_group y otra llamada num\_cases. Consulta la página sobre [Pivoteo de datos](#pivoting-data) para obtener más información sobre este proceso.

Ahora las primeras 50 filas de datos tienen este aspecto:

Cuando se pasan las dos variables a facet\_grid(), lo más fácil es utilizar la notación de fórmula (por ejemplo, x ~ y) donde x son filas e y son columnas. Aquí está el gráfico, utilizando facet\_grid() para mostrar los gráficos para cada combinación de las columnas age\_group y District.

### Ejes libres o fijos

Las escalas de los ejes que se muestran cuando se hacen las facetas son por defecto las mismas (fijas) en todas las facetas. Esto es útil para las comparaciones cruzadas, pero no siempre es apropiado.

Al utilizar facet\_wrap() o facet\_grid(), podemos añadir scales = "free\_y" para "liberar" o liberar los ejes y de los paneles para que se escalen adecuadamente a su subconjunto de datos. Esto es particularmente útil si los recuentos reales son pequeños para una de las subcategorías y las tendencias son difíciles de ver de otra manera. En lugar de "free\_y" también podemos escribir "free\_x" para hacer lo mismo con el eje x (por ejemplo, para las fechas) o "free" para ambos ejes. Ten en cuenta que en facet\_grid, las escalas de y serán las mismas para las facetas en la misma fila, y las escalas de x serán las mismas para las facetas en la misma columna.

Cuando se utiliza facet\_grid solamente, podemos añadir space = "free\_y" o space = "free\_x" para que la altura o la anchura real de la faceta sea ponderada a los valores de la figura en su interior. Esto sólo funciona si ya se ha aplicado la escala = "free" (y o x).

### Orden del nivel de los factores en las facetas

Consulta esta [entrada](https://juliasilge.com/blog/reorder-within/) sobre cómo reordenar los niveles de los factores dentro de las facetas.

## Almacenamiento de gráficos

### Guardar los gráficos

Por defecto, cuando se ejecuta un comando ggplot(), el gráfico se imprimirá en el panel de RStudio Plots. Sin embargo, también puede guardar el gráfico como un objeto utilizando el operador de asignación <- y dándole un nombre. Entonces no se imprimirá a menos que se ejecute el propio nombre del objeto. También puede imprimirlo envolviendo el nombre del trazado con print(), pero esto sólo es necesario en ciertas circunstancias, como si el trazado se crea dentro de un bucle for utilizado para imprimir múltiples trazados a la vez (véase la página [Iteración, bucles y listas](#iteration-loops-and-lists)).

### Modificación de gráficos guardados

Una cosa buena de **ggplot2** es que puedes definir un gráfico (como arriba), y luego añadirle capas empezando por su nombre. No tiene que repetir todos los comandos que crearon el gráfico original.

Por ejemplo, para modificar el gráfico age\_by\_wt que se definió anteriormente, para incluir una línea vertical a la edad de 50 años, sólo tendríamos que añadir un + y empezar a añadir capas adicionales al gráfico.

### Exportación de gráficos

La exportación de ggplots es fácil con la función ggsave() de **ggplot2**. Puedes funcionar de dos maneras, ya sea:

* Especifique el nombre del objeto de la gráfica, a continuación, la ruta del archivo y el nombre con la extensión
  + Por ejemplo: ggsave(my\_plot, here("plots", "my\_plot.png"))
* Ejecuta el comando con sólo una ruta de archivo, para guardar el último gráfico que se imprimió
  + Por ejemplo: ggsave(here("plots", "my\_plot.png"))

Puedes exportar como png, pdf, jpeg, tiff, bmp, svg, o varios otros tipos de archivos, especificando la extensión del archivo en la ruta del mismo.

También puede especificar los argumentos ancho =, alto = y unidades = (ya sea "in", "cm" o "mm"). También puede especificar dpi = con un número para la resolución del trazado (por ejemplo, 300). Consulta los detalles de la función introduciendo ?ggsave o leyendo la [documentación en línea](https://ggplot2.tidyverse.org/reference/ggsave.html).

Recuerde que puede utilizar la sintaxis here() para proporcionar la ruta de archivo deseada. Consulta la página de [importación y exportación](#import-and-export) para obtener más información.

## Etiquetas

Seguramente querrá añadir o ajustar las etiquetas del gráfico. Esto se hace más fácilmente dentro de la función labs() que se añade al gráfico con + al igual que los geoms.

Dentro de labs() puede proporcionar cadenas de caracteres a estos argumentos:

* x = e y = El título del eje x y del eje y (etiquetas)
* title = El título de la trama principal
* subtítulo = El subtítulo de la gráfica, en texto más pequeño debajo del título
* caption = El título del gráfico, en la parte inferior derecha por defecto

Aquí está un gráfico que hicimos antes, pero con etiquetas más bonitas:

Observe cómo en la asignación del título hemos utilizado str\_glue() del paquete **stringr** para implantar código R dinámico dentro del texto de la cadena. El título mostrará la fecha "Datos a partir de:" que refleja la fecha máxima de hospitalización en linelist. Lea más sobre esto en la página sobre [Caracteres y cadenas](#characters-and-strings).

Una nota sobre la especificación del título de la leyenda: No hay un argumento "título de la leyenda", ya que podría tener múltiples escalas en su leyenda. Dentro de labs(), puede escribir el argumento de la estética del gráfico utilizado para crear la leyenda, y proporcionar el título de esta manera. Por ejemplo, arriba asignamos color = edad para crear la leyenda. Por lo tanto, proporcionamos color = a labs() y asignamos el título de la leyenda deseado ("Edad" con A mayúscula). Si se crea la leyenda con aes(fill = COLUMN), entonces en labs() se escribiría fill = para ajustar el título de esa leyenda. La sección sobre escalas de color en la página [de consejos de ggplot](#ggplot-tips) proporciona más detalles sobre la edición de leyendas, y un enfoque alternativo utilizando las funciones scales\_().

## Temas

Una de las mejores partes de **ggplot2** es la cantidad de control que tienes sobre el gráfico - ¡puedes definir cualquier cosa! Como se mencionó anteriormente, el diseño del gráfico que no está relacionado con las formas/geometrías de los datos se ajustan dentro de la función theme(). Por ejemplo, el color de fondo del gráfico, la presencia/ausencia de líneas de cuadrícula, y la fuente/tamaño/color/alineación del texto (títulos, subtítulos, leyendas, texto de los ejes...). Estos ajustes pueden realizarse de dos maneras:

* Añade una función theme\_() [completa](https://ggplot2.tidyverse.org/reference/ggtheme.html) para realizar ajustes de barrido - incluye theme\_classic(), theme\_minimal(), theme\_dark(), theme\_light() theme\_grey(), theme\_bw() entre otras
* Ajuste cada pequeño aspecto de la trama individualmente dentro de theme()

### Temas completos

Como son bastante sencillas, demostraremos las funciones del tema completo a continuación y no las describiremos más aquí. Ten en cuenta que cualquier microajuste con theme() debe hacerse después de utilizar un tema completo.

Escríbalos con paréntesis vacíos.

### Modificar el tema

La función theme() puede tomar un gran número de argumentos, cada uno de los cuales edita un aspecto muy específico del gráfico. No hay manera de que podamos cubrir todos los argumentos, pero describiremos el patrón general para ellos y le mostraremos cómo encontrar el nombre del argumento que necesita. La sintaxis básica es esta:

1. Dentro de theme() escribe el nombre del argumento del elemento de la trama que quieres editar, como plot.title =
2. Proporcionar una función element\_() al argumento

* Lo más habitual es utilizar element\_text(), pero también element\_rect() para los colores de fondo del lienzo, o element\_blank() para eliminar los elementos de la trama

1. Dentro de la función element\_(), escriba las asignaciones de argumentos para realizar los ajustes finos que desee

Esa descripción era bastante abstracta, así que aquí hay algunos ejemplos.

El siguiente gráfico parece bastante tonto, pero sirve para mostrarle una variedad de formas en las que puede ajustar su gráfico.

* Comenzamos con el gráfico age\_by\_wt definido anteriormente y añadimos theme\_classic()
* Para realizar ajustes más finos, añadimos theme() e incluimos un argumento por cada elemento de la trama a ajustar

Puedes ser bueno organizar los argumentos en secciones lógicas. A continuación se describen algunos de los utilizados:

* legend.position = es único en el sentido de que acepta valores simples como "abajo", "arriba", "izquierda" y "derecha". Pero, por lo general, los argumentos relacionados con el texto requieren que se coloquen los detalles dentro de element\_text().
* Tamaño del título con element\_text(size = 30)
* La alineación horizontal del título con element\_text(hjust = 0) (de derecha a izquierda)
* El subtítulo está en cursiva con element\_text(face = "italic")

Aquí hay algunos argumentos de theme() especialmente comunes. Reconocerá algunos patrones, como añadir .x o .y para aplicar el cambio sólo a un eje.

| **argumento de theme()** | **Lo que ajusta** |
| --- | --- |
| plot.title = element\_text() | El título |
| plot.subtitle = element\_text() | El subtítulo |
| plot.caption = element\_text() | La leyenda (familia, cara, color, tamaño, ángulo, vjust, hjust...) |
| axis.title = element\_text() | Títulos de los ejes (tanto x como y) (tamaño, cara, ángulo, color...) |
| axis.title.x = element\_text() | Título del eje x solamente (use .y para el eje y solamente) |
| axis.text = element\_text() | Texto de los ejes (x e y) |
| axis.text.x = element\_text() | Texto del eje x solamente (use .y para el eje y solamente) |
| axis.ticks = element\_blank() | Eliminar las garrapatas del eje |
| axis.line = element\_line() | Líneas del eje (color, tamaño, tipo de línea: sólida, punteada, etc.) |
| strip.text = element\_text() | Texto de la tira de facetas (color, cara, tamaño, ángulo...) |
| strip.background = element\_rect() | tira de facetas (relleno, color, tamaño...) |

Pero ¡hay tantos argumentos temáticos! ¿Cómo podría recordarlos todos? No te preocupes, es imposible recordarlos todos. Por suerte, hay algunas herramientas que le ayudarán:

La documentación de **tidyverse** sobre la [modificación del tema](https://ggplot2.tidyverse.org/reference/theme.html), que tiene una lista completa.

**CONSEJO:** Ejecuta theme\_get() de **ggplot2** para imprimir una lista de los más de 90 argumentos de theme() en la consola.

**TIP:** Si alguna vez quieres eliminar un elemento de un gráfico, también puedes hacerlo a través de theme(). Basta con pasar element\_blank() a un argumento para que desaparezca por completo. Para las leyendas, establezca legend.position = "none".

## Colores

Consulta esta [sección sobre las escalas de color de la página de consejos de ggplot](#ggplot_tips_colors).

## Canalización en ****ggplot2****

Cuando se utilizan tuberías para limpiar y transformar los datos, es fácil pasar los datos transformados a ggplot().

Las tuberías que pasan el conjunto de datos de función a función pasarán a + una vez que se llame a la función ggplot(). Ten en cuenta que, en este caso, no es necesario especificar el argumento data =, ya que éste se define automáticamente como el conjunto de datos canalizado.

Así es como podría verse:

## Trazar datos continuos

A lo largo de esta página, ya has visto muchos ejemplos de trazado de datos continuos. Aquí los consolidamos brevemente y presentamos algunas variaciones.   
Las visualizaciones que aquí se cubren incluyen:

* Gráficos para una variable continua:
  + **Histograma**, un gráfico clásico para presentar la distribución de una variable continua.
  + **Gráfico de caja** (también llamado de caja y bigotes), para mostrar los percentiles 25, 50 y 75, los extremos de la cola de la distribución y los valores atípicos ([limitaciones importantes](https://www.data-to-viz.com/caveat/boxplot.html)).
  + **Gráfico de fluctuación**, para mostrar todos los valores como puntos que se "fluctúan" para que se puedan ver (en su mayoría) todos, incluso cuando dos tienen el mismo valor.
  + **Gráfico del violín**, muestra la distribución de una variable continua en función de la anchura simétrica del "violín".
  + Los **gráficos de Sina**, son una combinación de los gráficos de jitter y de violín, donde se muestran los puntos individuales pero con la forma simétrica de la distribución (a través del paquete **ggforce**).
* **Gráfico de dispersión** para dos variables continuas.
* **Gráficos de calor** para tres variables continuas (enlazado a la página de [gráficos de calor](#heat-plots))

### Histogramas

Los histogramas pueden parecerse a los gráficos de barras, pero son distintos porque miden la distribución de una variable continua. No hay espacios entre las "barras", y sólo se proporciona una columna a geom\_histogram().

A continuación se muestra el código para generar **histogramas**, que agrupan datos continuos en rangos y se muestran en barras adyacentes de altura variable. Esto se hace utilizando geom\_histogram(). Consulta la [sección "Diagrama de barras"](#ggplot_basics_bars) de la página de fundamentos de ggplot para entender la diferencia entre geom\_histograma(), geom\_bar() y geom\_col().

Vamos a mostrar la distribución de las edades de los casos. Dentro de mapping = aes() especifique la columna de la que quiere ver la distribución. Puedes asignar esta columna al eje x o al eje y.

Las filas serán asignadas a "bins" basados en su edad numérica, y estos bins serán representados gráficamente por barras. Si se especifica un número de bins con la estética de gráfico bins =, los puntos de ruptura se espacian uniformemente entre los valores mínimos y máximos del histograma. Si no se especifica bins =, se adivinará un número apropiado de bins y se mostrará este mensaje después del gráfico:

## `stat\_bin()` usando `bins = 30`. Elija un valor mejor con `binwidth`.

Si no quiere especificar un número de bins a bins =, puede especificar alternativamente binwidth = en las unidades del eje. Damos algunos ejemplos que muestran diferentes bins y anchos de bins:

Para obtener proporciones suavizadas, puede utilizar geom\_density():

Para obtener un histograma "apilado" (de una columna continua de datos), puede hacer una de las siguientes cosas:

1. Utiliza geom\_histogram() con el argumento fill = dentro de aes() y asignado a la columna de agrupación, o
2. Utiliza geom\_freqpoly(), que probablemente sea más fácil de leer (aún puedes establecer binwidth =)
3. Para ver las proporciones de todos los valores, establezca el y = after\_stat(density) (utiliza esta sintaxis exactamente - no se ha modificado para sus datos). Nota: estas proporciones se mostrarán por grupo.

Cada uno se muestra a continuación (\*nótese el uso de color = frente a relleno = en cada uno):

Si quieres divertirte un poco, prueba con geom\_density\_ridges del paquete **ggridges** ([vignette aquí](https://cran.r-project.org/web/packages/ggridges/vignettes/introduction.html).

Lea más en detalle sobre los histogramas en la [página de](https://ggplot2.tidyverse.org/reference/geom_histogram.html) **tidyverse** [sobre geom\_histogram()](https://ggplot2.tidyverse.org/reference/geom_histogram.html).

### Gráficos de caja

Los gráficos de caja son habituales, pero tienen importantes limitaciones. Pueden ocultar la distribución real - por ejemplo, una distribución bimodal. Consulta esta [galería de gráficos de R](https://www.r-graph-gallery.com/boxplot.html) y este [artículo sobre la conversión de datos en imágenes](https://www.data-to-viz.com/caveat/boxplot.html) para obtener más detalles. Sin embargo, muestran muy bien el rango intercuartil y los valores atípicos, por lo que pueden superponerse a otros tipos de gráficos que muestran la distribución con más detalle.

A continuación le recordamos los distintos componentes de un boxplot:

Cuando se utiliza geom\_boxplot() para crear un gráfico de caja, generalmente se asigna sólo un eje (x o y) dentro de aes(). El eje especificado determina si los gráficos son horizontales o verticales.

En la mayoría de los geoms, se crea un gráfico por grupo asignando una estética como color = o relleno = a una columna dentro de aes(). Sin embargo, en el caso de los gráficos de caja, esto se consigue asignando la columna de agrupación al eje no asignado (x o y). A continuación se muestra el código para un diagrama de caja de todos los valores de edad en el conjunto de datos, y en segundo lugar el código para mostrar un diagrama de caja para cada género (no ausente) en el conjunto de datos. Ten en cuenta que los valores NA (que faltan) aparecerán como un gráfico de caja separado a menos que se eliminen. En este ejemplo, también hemos establecido el relleno del resultado de la columna para que cada gráfico tenga un color diferente, pero esto no es necesario.

Para ver el código para añadir un gráfico de caja a los bordes de un gráfico de dispersión (gráficos "marginales"), consulta la página [ggplot tips](#ggplot-tips).

### Gráficos de violín, jitter y sina

A continuación se muestra el código para crear gráficos de **violín** (geom\_violin) y **gráficos de fluctuación** (geom\_jitter) para mostrar distribuciones. Se puede especificar que el relleno o el color también estén determinados por los datos, insertando estas opciones dentro de aes().

Puedes combinar los dos usando la función geom\_sina() del paquete **ggforce**. La sina traza los puntos de jitter en la forma del gráfico de violín. Cuando se superpone al gráfico de violín (ajustando las transparencias) puede ser más fácil de interpretar visualmente.

### Dos variables continuas

Siguiendo una sintaxis similar, geom\_point() le permitirá trazar dos variables continuas entre sí en un **gráfico de dispersión**. Esto es útil para mostrar los valores reales en lugar de sus distribuciones. En (A) se muestra un gráfico de dispersión básico de la edad frente al peso. En (B) volvemos a utilizar facet\_grid() para mostrar la relación entre dos variables continuas en linelist.

### Tres variables continuas

Puedes mostrar tres variables continuas utilizando el argumento fill = para crear un gráfico de calor. El color de cada "celda" reflejará el valor de la tercera columna continua de datos. Consulta la página de [consejos de ggplot](#ggplot-tips) y la página de [gráficos de calor](#heat-plots) para obtener más detalles y varios ejemplos.

Hay formas de hacer gráficos en 3D en R, pero para la epidemiología aplicada suelen ser difíciles de interpretar y, por tanto, menos útiles para la toma de decisiones.

## Trazar datos categóricos

Los datos categóricos pueden ser valores de carácter, pueden ser lógicos (TRUE/FALSE), o factores (ver la página de [Factores](#factors)).

### Preparación

#### Estructura de datos

Lo primero que hay que entender sobre sus datos categóricos es si existen como observaciones brutas, como una lista de casos, o como un dataframe de resumen o agregado que contiene recuentos o proporciones. El estado de sus datos afectará a la función de trazado que utilice:

* Si sus datos son observaciones en bruto con una fila por observación, es probable que utiliza geom\_bar()
* Si sus datos ya están agregados en recuentos o proporciones, es probable que utiliza geom\_col()

#### Tipo de columna y ordenación de valores

A continuación, examine el tipo de las columnas que desea trazar. Miramos a hospital, primero con class() de , y con tabyl() de **janitor**.

Podemos ver que los valores dentro son caracteres, ya que son nombres de hospitales, y por defecto están ordenados alfabéticamente. Hay valores "otros" y "desaparecidos", que preferiríamos que fueran las últimas subcategorías al presentar los desgloses. Así que cambiamos esta columna por un factor y la reordenamos. Esto se trata con más detalle en la página de [Factores](#factors).

### geom\_bar()

Utiliza geom\_bar() si desea que la altura de las barras (o la altura de los componentes de las barras apiladas) refleje el número de filas relevantes de los datos. Estas barras tendrán huecos entre ellas, a menos que se ajuste la estética de width = plot.

* Proporcione sólo una asignación de columna de eje (normalmente el eje x). Si proporciona x e y, obtendrá Error: stat\_count() sólo puede tener una estética x o y.
* Puedes crear barras apiladas añadiendo una asignación de relleno = columna dentro de mapping = aes()
* El eje opuesto se llamará por defecto "recuento", ya que representa el número de filas

A continuación, hemos asignado el resultado al eje Y, pero podría estar fácilmente en el eje X. Si tiene valores de caracteres más largos, a veces puede parecer mejor invertir las barras de lado y poner la leyenda en la parte inferior. Esto puede afectar a la forma en que se ordenan los niveles de los factores - en este caso los invertimos con fct\_rev() para poner los faltantes y otros en la parte inferior.

### geom\_col()

Utiliza geom\_col() si desea que la altura de las barras (o la altura de los componentes de las barras apiladas) refleje valores precalculados que existen en los datos. A menudo, se trata de recuentos sumarios o "agregados", o de proporciones.

Proporcione las asignaciones de columna para ambos ejes a geom\_col(). Normalmente la columna del eje x es discreta y la del eje y es numérica.

Digamos que tenemos los resultados de este conjunto de datos:

A continuación se muestra un código que utiliza geom\_col para crear gráficos de barras sencillos que muestren la distribución de los resultados de los pacientes con ébola. Con geom\_col, es necesario especificar tanto x como y. Aquí x es la variable categórica a lo largo del eje x, e y es la proporción de la columna de proporciones generada.

Para mostrar los desgloses por hospital, necesitaríamos que nuestra tabla contuviera más información, y que estuviera en formato "largo". Creamos esta tabla con las frecuencias de las categorías combinadas resultado y hospital (véase la página de [Agrupar datos](#grouping-data) para obtener consejos de agrupación).

A continuación, creamos el ggplot con algún formato añadido:

* **Cambio de eje**: Intercambiamos los ejes con coord\_flip() para poder leer los nombres de los hospitales.
* **Columnas de lado a lado**: Se ha añadido un argumento de posición = "esquivar" para que las barras de muerte y recuperación se presenten una al lado de la otra en lugar de apiladas. Ten en cuenta que las barras apiladas son las predeterminadas.
* **Ancho de columna**: Se especifica "ancho", por lo que las columnas son la mitad del ancho posible.
* **Orden de las columnas**: Se ha invertido el orden de las categorías en el eje y para que "Otros" y "Falta" estén en la parte inferior, con scale\_x\_discrete(limits=rev). Ten en cuenta que usamos eso en lugar de scale\_y\_discrete porque hospital se indica en el argumento x de aes(), aunque visualmente esté en el eje y. Hacemos esto porque Ggplot parece presentar las categorías al revés a menos que le digamos que no lo haga.
* **Otros detalles**: Etiquetas/títulos y colores añadidos dentro de labs y scale\_fill\_color respectivamente.

Ten en cuenta que las proporciones son binarias, por lo que podemos preferir omitir "recuperar" y mostrar sólo la proporción que murió. Esto es sólo a título ilustrativo.

Si se utiliza geom\_col() con datos de fechas (por ejemplo, una epicurva a partir de datos agregados) - querrá ajustar el argumento width = para eliminar las líneas de "hueco" entre las barras. Si se utilizan datos diarios, ajuste el ancho = 1. Si se trata de datos semanales, la anchura es de 7. Los meses no son posibles porque cada mes tiene un número diferente de días.

### geom\_histograma()

Los histogramas pueden parecerse a los gráficos de barras, pero son distintos porque miden la distribución de una variable continua. No hay espacios entre las "barras" y sólo se proporciona una columna a geom\_histogram(). Hay argumentos específicos para los histogramas como bin\_width = y breaks = para especificar cómo se deben dividir los datos. La sección anterior sobre datos continuos y la página sobre [curvas epidémicas](#epidemic-curves) proporcionan detalles adicionales.

## Recursos

Hay una gran cantidad de ayuda en línea, especialmente con ggplot. Ver:

* [Hoja de trucos de ggplot2](http://r-statistics.co/ggplot2-cheatsheet.html)
* [otra hoja de trucos](https://biostats.w.uib.no/the-ggplot2-cheat-sheet-by-rstudio/)
* [página de fundamentos de tidyverse ggplot](https://ggplot2.tidyverse.org/reference/)
* [trazado de variables continuas](http://www.sthda.com/english/articles/32-r-graphics-essentials/131-plot-two-continuous-variables-scatter-graph-and-alternatives/)
* Páginas de R for Data Science sobre [visualización de datos](https://r4ds.had.co.nz/data-visualisation.html)
* [gráficos para la comunicación](https://r4ds.had.co.nz/graphics-for-communication.html)