

CACCS: Secuencia de taller de RStudio – Parte 2

Rashid C.J. Marcano Rivera

4 de oct. de 2024

Contents

Visualización de datos: porqués	1
ggplot2	4
Componentes de un gráfico	5
Datos del Covid	17
Datos del censo	20
Recapitulando	29
Qué aprendimos en este taller inicial	29
Continuamos el próximo viernes	29
Visualización de datos	

Este taller está basado en viñetas de RStudio con tidyverse (en específico `ggplot2`) así como elementos y ejemplos del libro de Rafael Irizarry disponible aquí o de manera similar a los html producidos en este taller y más actualizada aquí. Finalmente la parte del censo proviene en parte de código ajustado del libro Public Policy Analytics: Code & Context for Data Science in Government, por Ken Steif.

Si aún no has instalado R, está aquí. Acto seguido, baja RStudio. Puedes también ir a la nube en Posit Cloud.

Visualización de datos: porqués

Ver números y cadenas de caracteres que forman un conjunto de datos puede ser interesante, o no, pero normalmente no tiene tanta utilidad. Por ejemplo

```
library(wooldridge)  
  
data(wage1)  
head(wage1)
```

```
##   wage educ exper tenure nonwhite female married numdep smsa northcen south  
## 1 3.10    11     2      0       0     1       0      2     1       0     0  
## 2 3.24    12    22      2       0     1       1      3     1       0     0
```

```

## 3 3.00 11 2 0 0 0 2 0 0 0
## 4 6.00 8 44 28 0 0 1 0 1 0 0
## 5 5.30 12 7 2 0 0 1 1 0 0 0
## 6 8.75 16 9 8 0 0 1 0 1 0 0
##   west construc ndurman trcommu trade services profserv profocc clerocc
## 1 1 0 0 0 0 0 0 0 0 0
## 2 1 0 0 0 0 1 0 0 0 0
## 3 1 0 0 0 1 0 0 0 0 0
## 4 1 0 0 0 0 0 0 0 0 1
## 5 1 0 0 0 0 0 0 0 0 0
## 6 1 0 0 0 0 0 1 1 1 0
##   servocc lwage expersq tenursq
## 1 0 1.131402 4 0
## 2 1 1.175573 484 4
## 3 0 1.098612 4 0
## 4 0 1.791759 1936 784
## 5 0 1.667707 49 4
## 6 0 2.169054 81 64

```

¿Qué aprendemos de ver estos datos así? ¿Podemos rápidamente determinar a si años de educación se traducen a mayores ingresos? ¿Podemos determinar si afecta en algo la relación marital? Para muchos humanos, es difícil extraer información con meramente mirar a números sin contexto adicional. Pero podríamos ver algo en este gráfico

Lo mismo podríamos hacer con los datos que trabajamos la semana pasada de homicidios con armas de fuego en EEUU:

```

library(dslabs)
library(tidyverse)

## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr     1.1.4    v readr     2.1.5
## vforcats   1.0.0    v stringr   1.5.1
## v ggplot2   3.5.1    v tibble    3.2.1
## v lubridate 1.9.3    v tidyrr    1.3.1
## v purrr    1.0.2
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()   masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

data("murders")
head(murders)

##      state abb region population total
## 1 Alabama AL  South  4779736  135
## 2 Alaska AK  West   710231   19
## 3 Arizona AZ  West  6392017  232
## 4 Arkansas AR  South  2915918   93
## 5 California CA  West  37253956 1257
## 6 Colorado CO  West  5029196   65

```

No podemos determinar con facilidad a qué estado le toca la población más grande o pequeña, y si existe alguna relación entre el tamaño de población y el total de asesinatos, o de cómo varían las tasas de asesinatos

Relación entre educación y salario

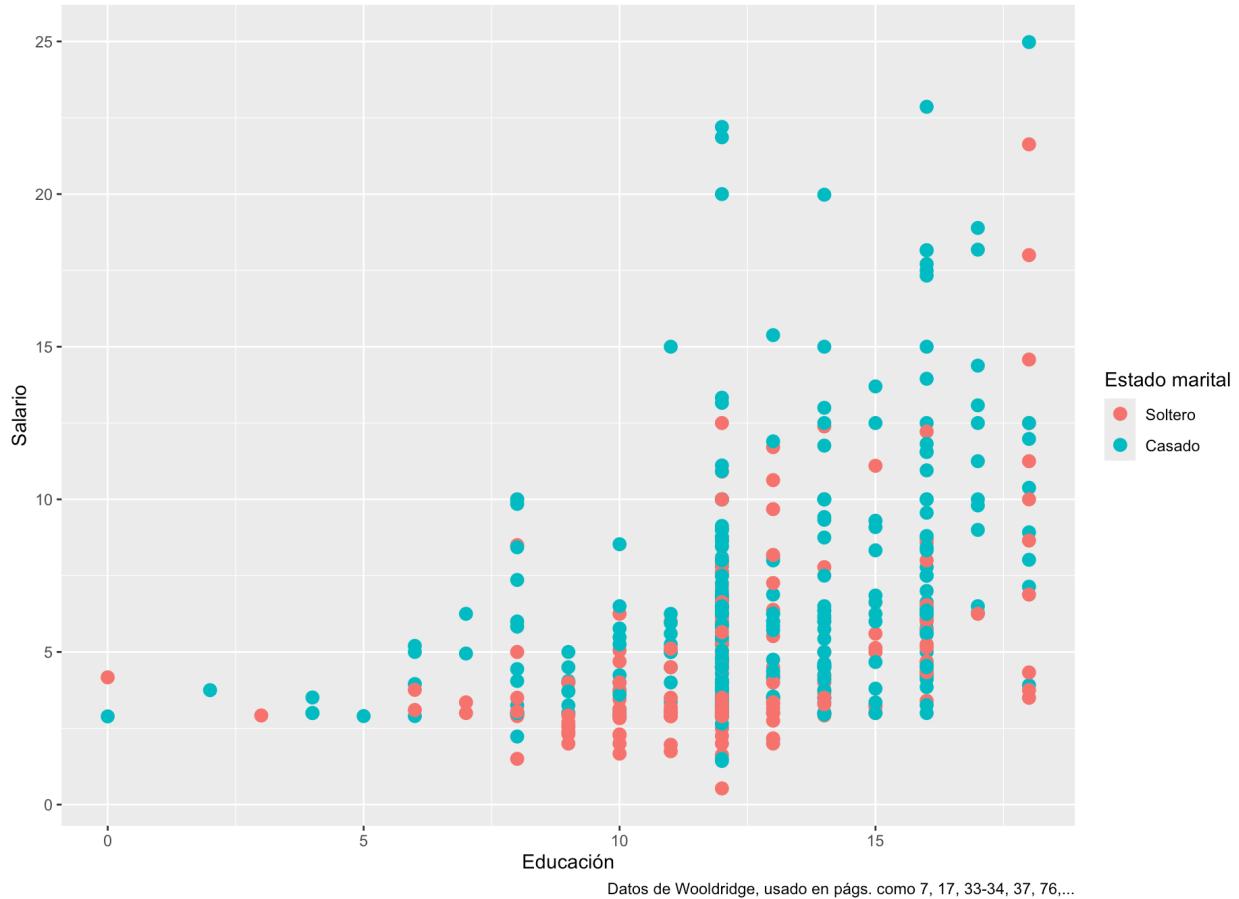


Figure 1: Meta 1

por regiones de estados en la federación estadounidense. Sin embargo, eso puede responderse sin muchas palabras con el próximo gráfico.

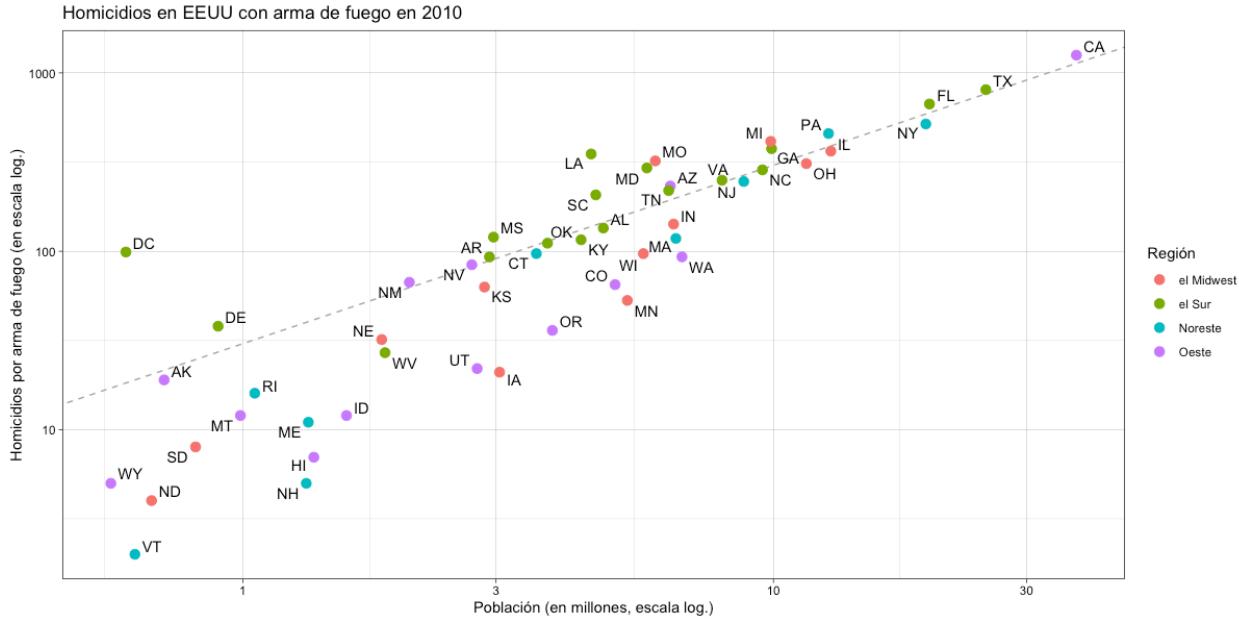


Figure 2: Gráfico de meta 2.

Una imagen vale más que mil palabras dice el dicho. Sin adentrarnos a la inferencia estadística (que cubriremos en la tercera y última semana de esta secuencia de talleres), hemos podido comunicar relaciones y hallazgos en datos. A veces puede ser este ejercicio uno tan convincente que no requiera análisis subsiguientes.

Vivimos en una era de creciente disponibilidad de conjuntos de datos informativos y de herramientas de software, con lo cual el uso de visualizaciones ha aumentado en diversos espacios: académicos, gubernamentales, organizaciones sociales, prensa, e industrias varias.

En R, una de las principales maneras con la que trabajaremos estos análisis visuales es de la mano de `ggplot2`.

Esta es una secuencia de tres semanas, y en esta lección continuamos con lo aprendido la semana anterior, donde terminamos con visualizaciones sencillas y con el uso de `tidyverse` para manejar datos. Tenemos la meta hoy de que al culminar las segundas dos horas de esta secuencia podamos:

1. Entender cómo usar la gramática de gráficas
2. Entender cómo usar `ggplot2`, continuando con lo aprendido de `tidyverse` de la semana pasada.
3. Entender cómo utilizar en varias maneras `tidycensus` para generar mapas.

La próxima sesión, del 18 de octubre, entraremos más en *wrangling* de datos, así como en la inferencia estadística, con introducciones en R sobre modelos lineales, jerárquicos y longitudinales.

ggplot2

R ofrece varias opciones para graficar, siendo útiles las capacidades incluidas en su instalación básica. Además, existen paquetes como `grid` y `lattice`. Sin embargo, en este libro se ha optado por usar `ggplot2`,

ya que permite a los principiantes crear gráficos complejos y estéticos mediante una sintaxis intuitiva y fácil de recordar, dividiendo los gráficos en componentes básicos.

`ggplot2` destaca por su uso de una *gramática de gráficos*, que simplifica el proceso de creación de gráficos. Al aprender unos pocos componentes esenciales de esta gramática, los usuarios pueden generar una amplia variedad de gráficos con facilidad. Además, su comportamiento por defecto está diseñado para producir resultados agradables y funcionales con código conciso y legible, lo que facilita su uso por principiantes. Como factor limitante está el que está diseñado para trabajar con tablas de formato *tidy* (con filas con observaciones y columnas conteniendo variables), pero un número sustancial de conjuntos de datos se trabajan en ese formato, o pueden convertirse como tal.

Componentes de un gráfico

Hoy construiremos varios tipos de gráficos como los que vimos arriba, así como mapas informativos. Antes que todo eso, vale señalar que los gráficos se dividen en tres componentes principales. Usaré otro gráfico que creara en 2020 durante la pandemia para ejemplificar estos componentes.

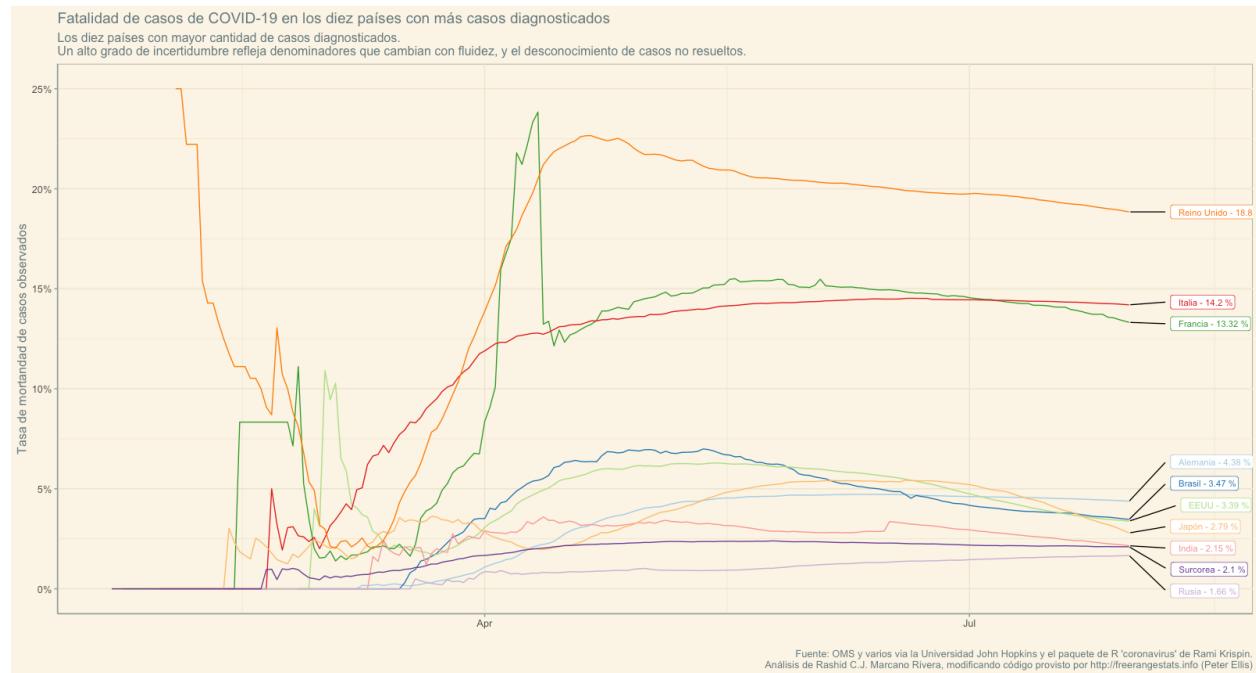


Figure 3: Datos para ejemplificar

- **Datos:**

- Estoy pasando al gráfico un conjunto de datos que corté sobre fatalidad de casos de COVID-19 hasta el 31 de julio de 2020 (los datos continúan hasta 2023).
- Este es el componente de *datos* del gráfico.

- **Geometrías:**

- El gráfico es uno de líneas, útil para varias series de tiempo. Este componente es una geometría. Otras geometrías posibles son gráficos de dispersión, histogramas, diagramas de barras, densidades suavizadas, y diagrama de cajas, entre otros.

- **Mapeo estético:**

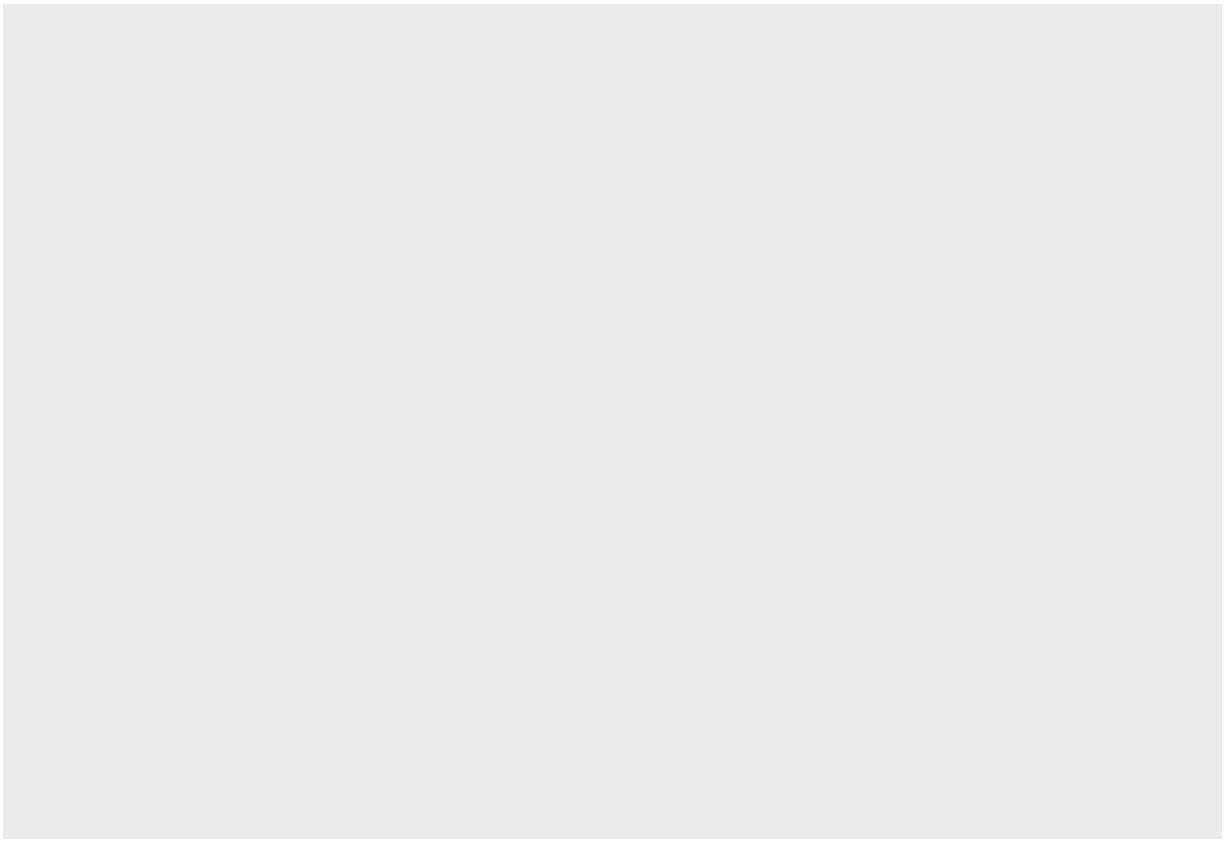
- El gráfico usa señales visuales para representar en el lienzo vacío con capas distintos tipos de información provista en el conjunto de datos:
 - * Posiciones en el eje de x (tiempo)
 - * Posiciones en el eje de y (tasas de mortandad observadas)
 - * Color (asignado por país)
- Cada línea representa la información de un país para una serie de fechas. Estos se aclaran con una etiqueta para aclararnos esa relación de línea-color-país.
- El mapeo estético depende de la geometría utilizada.

- **Observaciones adicionales:**

- Ejes x e y definidos por el rango de los datos y ambos en escalas logarítmicas.
- El gráfico incluye etiquetas, un título, etiquetas de variables, nota al calce, y el tema utilizado es uno solarizado que pareciera no muy distante al utilizado por el periódico “Financial Times”.
- Volveremos luego a estos datos para construir esta imagen si nos da tiempo en el taller, y si no, tendrán disponible el cómo hacerlo para referencia.

Manteniéndonos cerca de los datos utilizados en la semana pasada, construiremos por *capas* la información que va en el gráfico.

```
murders |> ggplot()
```



El primer paso de crear un gráfico de ggplot es asignar los datos a un lienzo vacío. Esto no significa poblar el lienzo con esos datos, sino pasarle al programa la información inicial, como un pintor que selecciona el tema que usará para la pintura que visualiza en su mente. Esto lo hicimos al pasar el *pipe* (`|>` o `%>%`) los datos a ggplot, y lo que ocurrió fue que al no darle más información (capas, como la pintura en un lienzo), nos quedó un recuadro gris enmarcado por un borde blanco.

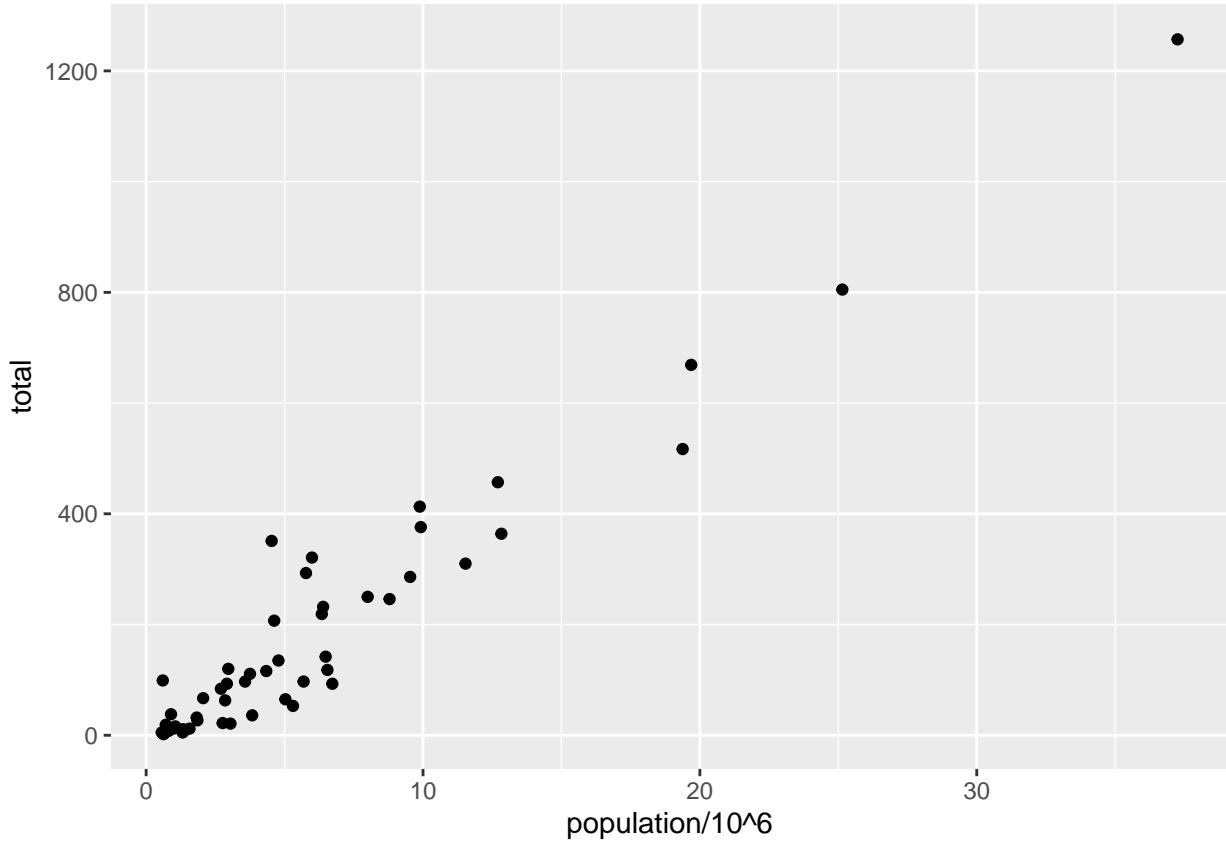
En ggplot, la información entonces se suministra al lienzo por capas, y se le pueden añadir adicionales. Esto tomará la forma de código siguiente:

```
DATOS |> ggplot() |> CAPA 1 |> CAPA 2 |> ... |> CAPA N
```

Usualmente, la primera capa que añadimos define la geometría. Si queremos hacer un diagrama de dispersión, ¿qué geometría deberíamos utilizar?

Si vemos la hoja de referencia (en la carpeta del taller 2, o accesible en esta página: <https://github.com/rstudio/cheatsheets/blob/main/data-visualization.pdf>), vemos que la función utilizada para crear gráficos con esta geometría puntillista es `geom_point`.

```
murders %>%
  ggplot() +
  geom_point(aes(x = population/10^6, y = total))
```

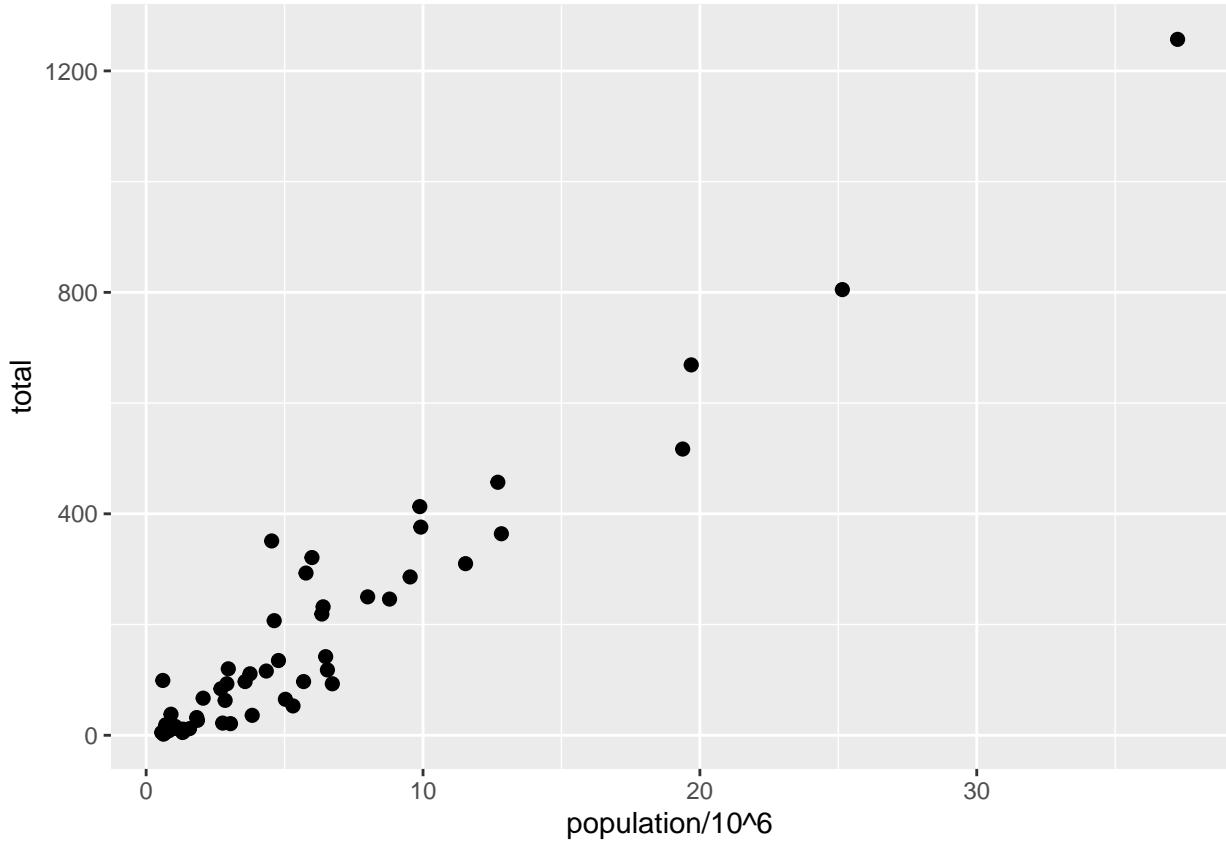


En este caso ya hemos dado un paso adicional y añadido una primera capa en esta obra: le indicamos que queremos una capa que tenga una geometría de puntos, y un mapeo estético que toma las coordenadas en un plano cartesiano donde el eje de x quedó definido como `population/10^6`, la población de los estados o Washington D.C., en millones; el eje de y quedó definido entonces como el total de asesinatos con armas de fuego. En el mapeo estético entonces los puntos quedan asignados a esas coordenadas. De esta manera, las distancias entre puntos, así como otras características que queramos añadir, quedan expresadas. Esto se da a través de la función `aes`. Esta será de las funciones que más usen al graficar.

Noten que hasta ahora hemos trabajado este lienzo sin guardarlo como objeto. Si bien esto puede funcionar bien, es posible que queramos guardar nuestro progreso y seguir añadiendo capas adicionales. En este caso, al ejecutar el comando y guardarlo en objeto, el programa no nos dará automáticamente una actualización del gráfico; tendremos que llamar al objeto para que aparezca:

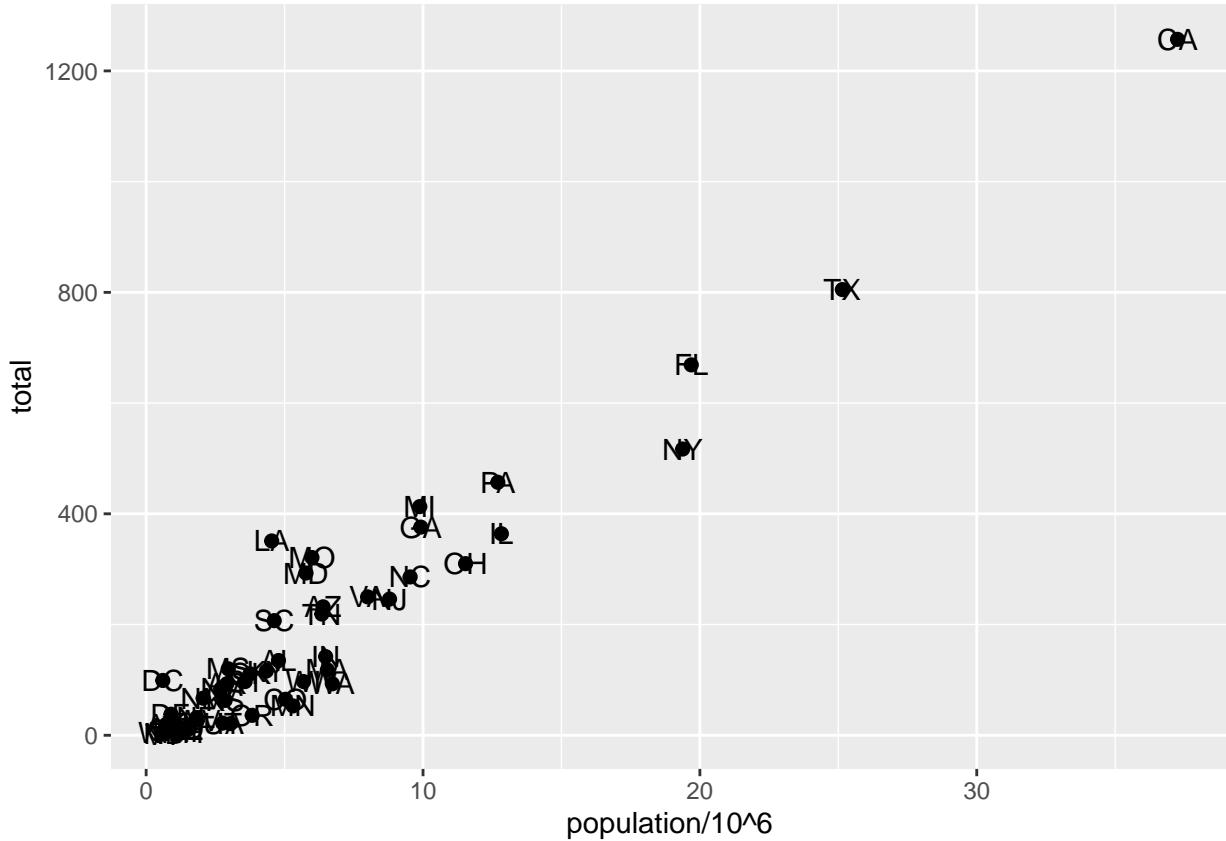
```
p<-murders %>%
  ggplot() +
  geom_point(aes(x = population/10^6, y = total), size=2)
```

p



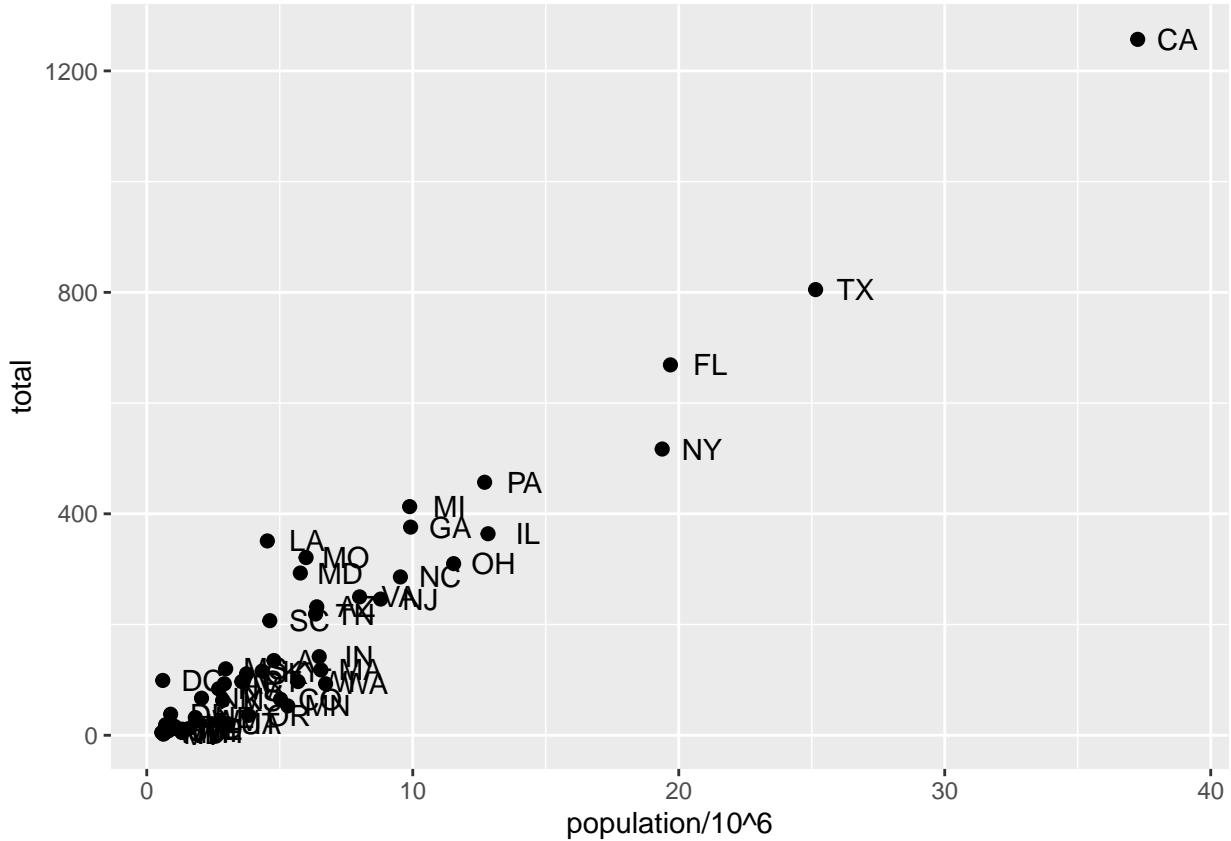
Tenemos entonces nuestro gráfico de dispersión inicial. Quizás no es tan informativo pero vemos una serie de puntos con el mapeo estético inicial. Noten que podríamos quitar la `x=` y la `y=` y no pasaría nada, ya que en ausencia de esta especificación, el programa entiende por defecto que lo primero que se le asigna es la información del eje horizontal, y en segundo orden el vertical:

```
p+
  geom_text(aes(population/10^6, total, label = abb))
```



Aquí he añadido una geometría nueva: texto. Hay dos geometrías para esto, `geom_label` y `geom_text`, uno con el texto enmarcado en un recuadro y el segundo sin ello. Ya que cada punto (cada jurisdicción que es realmente parte de los Estados Unidos de América en este caso) tiene una etiqueta, necesitamos un mapeo estético para hacer la conexión entre los puntos y las etiquetas, así que se le asignó el mismo tal que el texto cayera exactamente en la misma coordenada que el punto. Pero esto se puede corregir:

```
p+  
  geom_text(aes(population/10^6, total, label = abb), nudge_x = 1.5)
```

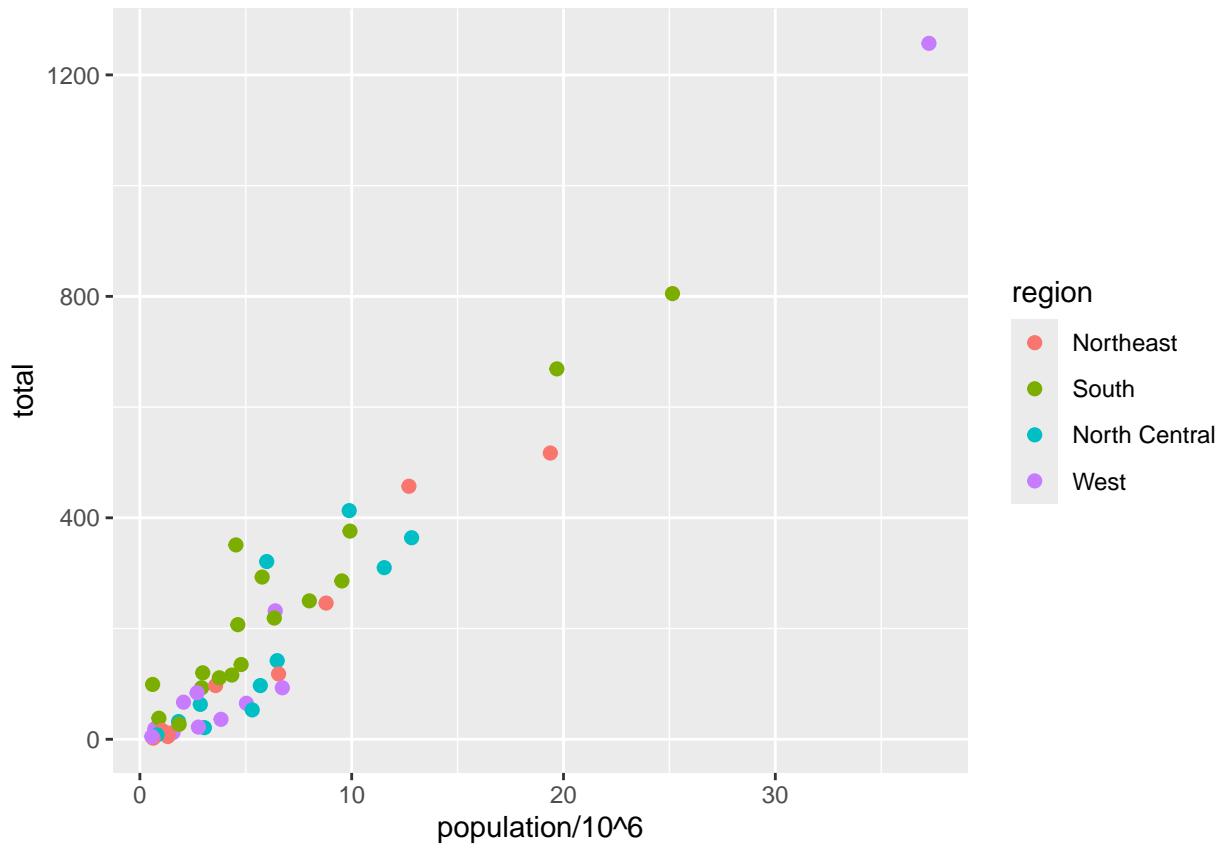


En este caso hemos empujado a través del eje de equis la etiqueta de texto con un valor numérico de 1.5. Valores mayores aumentarían la distancia del texto y el punto, mientras que menores harían lo contrario.

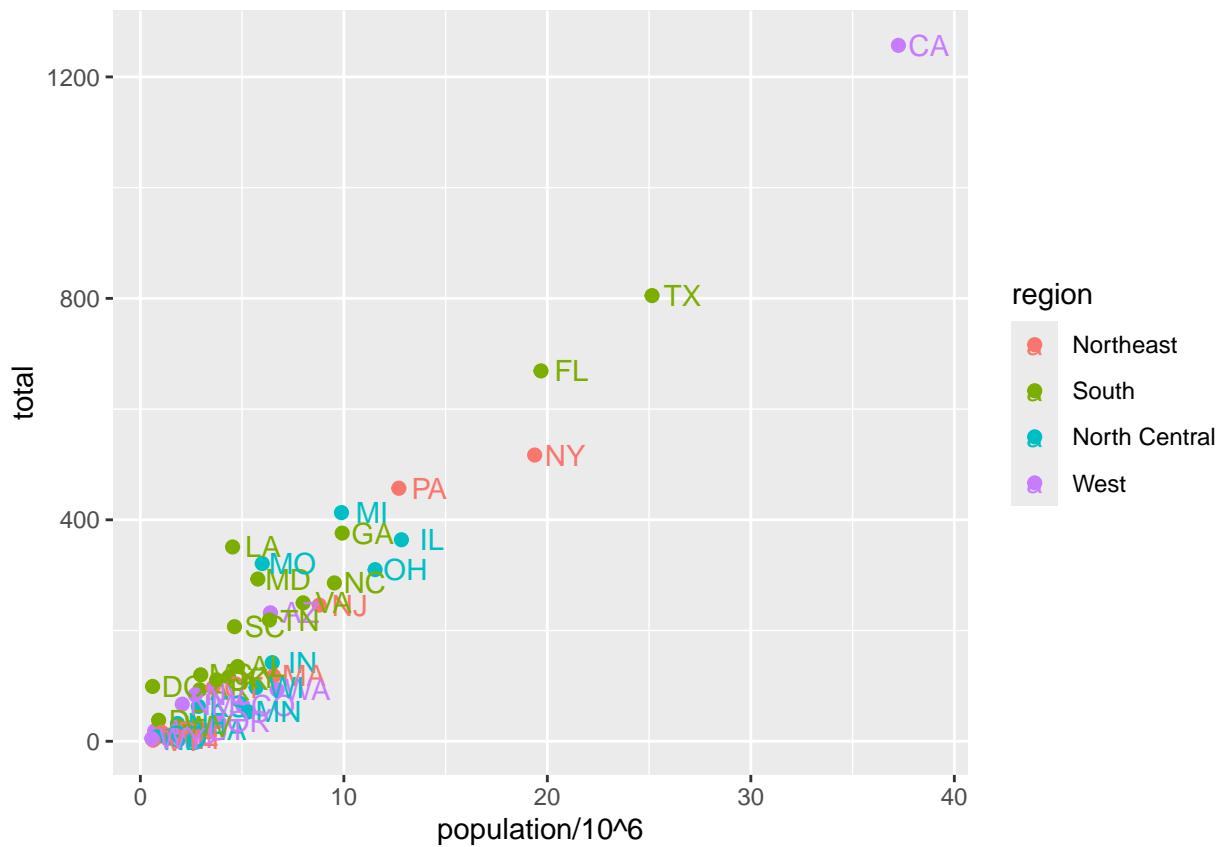
Ahora podremos seguir con una complicación en el mapeo estético: queremos añadir una capa de color al lienzo que represente regiones de estas jurisdicciones. Esto se hace al añadir la opción de `colour` y dándole una variable, en este caso `region`.

```
# Crear el gráfico base con puntos coloreados por región
p <- murders %>%
  ggplot() +
  geom_point(aes(x = population/10^6, y = total, colour = region), size = 2)

# Mostrar el gráfico
p
```



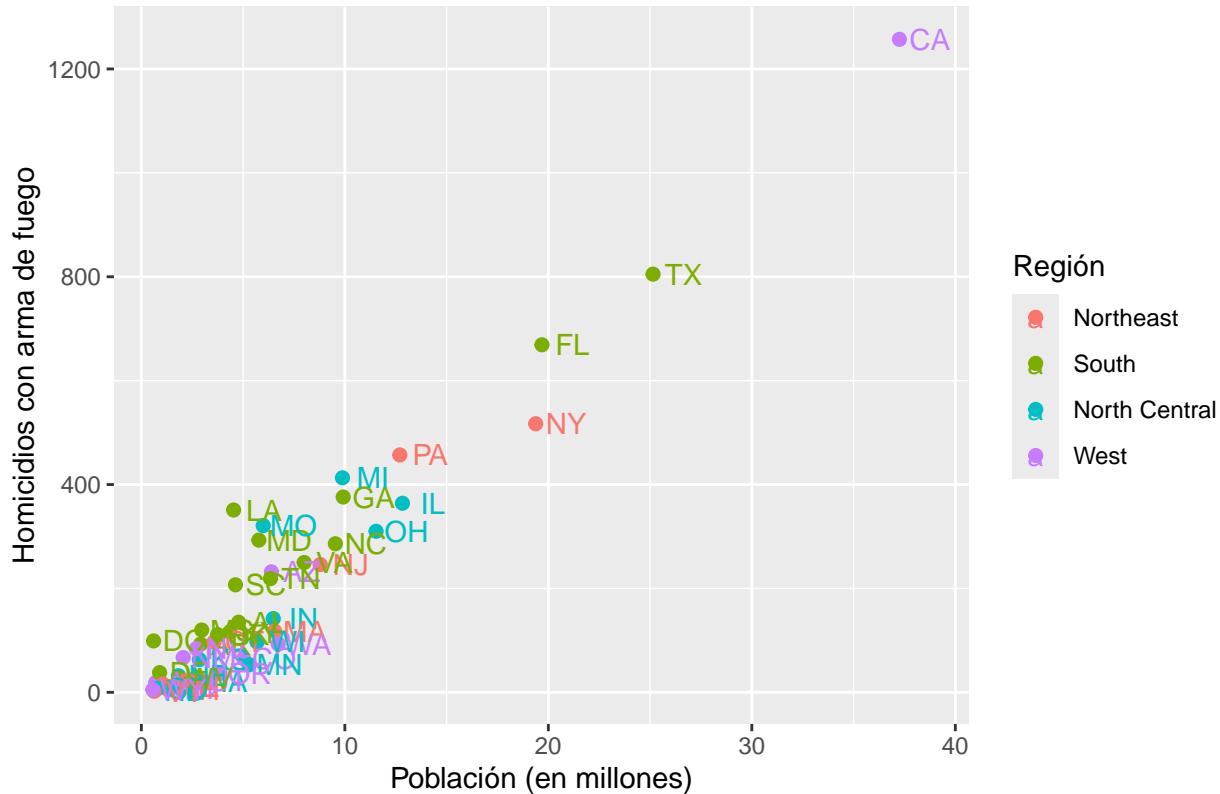
```
# Añadir etiquetas desplazadas en el eje x (con color por región)
p<-p +
  geom_text(aes(x = population/10^6, y = total, label = abb, colour = region), nudge_x = 1.5)
p
```



#demosle más información a la gráfica de dispersión.

```
p+labs(
  x = "Población (en millones)",      # Cambia el nombre del eje x
  y = "Homicidios con arma de fuego", # Cambia el nombre del eje y
  colour = "Región",                  # Cambia el nombre de la leyenda de color
  title = "Homicidios con arma de fuego en EEUU, 2010" # Título del gráfico
)
```

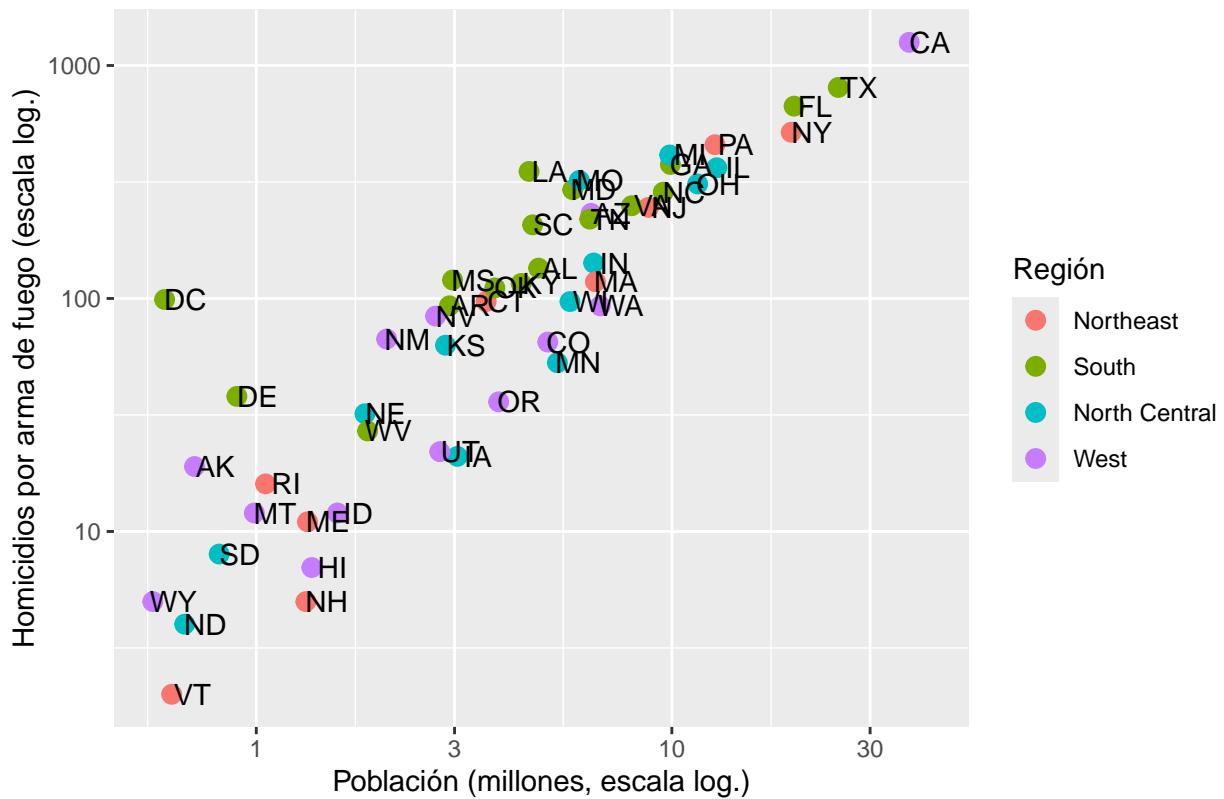
Homicidios con arma de fuego en EEUU, 2010



Hasta ahora hemos ido añadiendo capas pero notamos que tenemos una gran concentración de puntos en la parte inferior izquierda del lienzo: la mayoría de las jurisdicciones tienen menos de 10 millones de habitantes y menos de 400 homicidios. Esto hace leer e interpretar lo que sucede en para estos casos difícil. Podríamos entonces representar el gráfico con una transformación logarítmica al re-escalar con `scale_x_continuous` y `scale_y_continuous`:

```
p2<-murders %>%
  ggplot() +
  geom_point(aes(x = population/10^6, y = total, colour = region), size = 3)
p2<- p2 + geom_text(aes(x = population/10^6, y = total, label = abb), nudge_x = 0.05)
p2<-p2+scale_x_continuous(trans = "log10")+
  scale_y_continuous(transform = "log10")
p2<-p2+
  labs(
    x = "Población (millones, escala log.)", # Cambia el nombre del eje x
    y = "Homicidios por arma de fuego (escala log.)", # Cambia el nombre del eje y
    colour = "Región", # Cambia el nombre de la leyenda de color
    title = "Homicidios por arma de fuego vs población, por región" # Título del gráfico
  )
p2
```

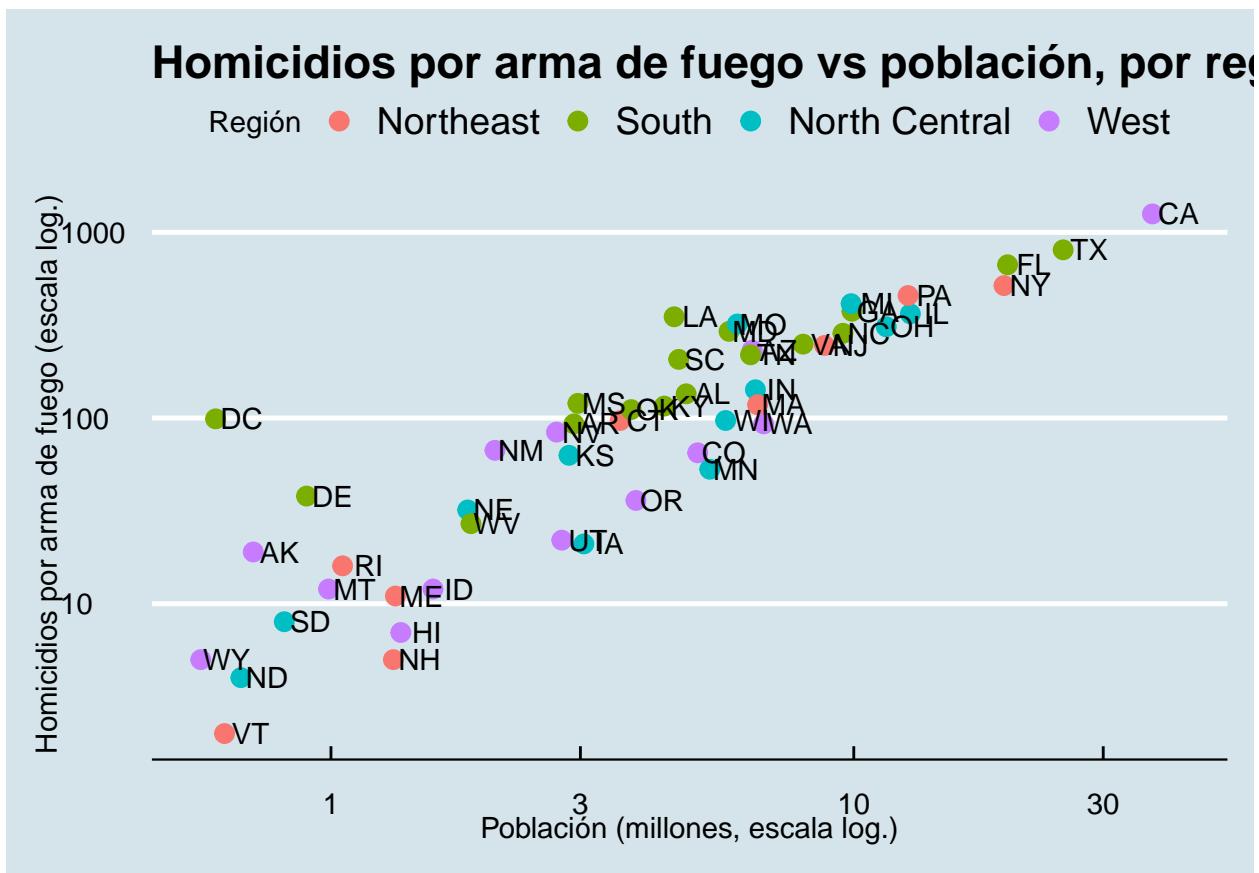
Homicidios por arma de fuego vs población, por región



Podríamos añadir una capa temática, usando el paquete `ggthemes`, que tiene un catálogo estético variado, que recomiendo verifiquen a través de <https://yutannihilation.github.io/allYourFigureAreBelongToUs/ggthemes/>.

En este caso le añadiré una visualización al estilo del semanario británico *The Economist*.

```
library(ggthemes)
p2 + theme_economist()
```



Normalmente queremos añadir formas o anotaciones a las figuras que no se derivan directamente del mapeo estético; algunos ejemplos incluyen etiquetas, cuadros, áreas sombreadas y líneas. Si queremos añadir una línea que represente la tasa promedio de asesinatos en todos los Estados Unidos, tendremos que determinarlo aparte con la ayuda de `dplyr` (parte de `tidyverse`), y tendremos que hacer la transformación adecuada también (logarítmica):

```
library(ggthemes)
library(ggrepel)

## Warning: package 'ggrepel' was built under R version 4.3.3

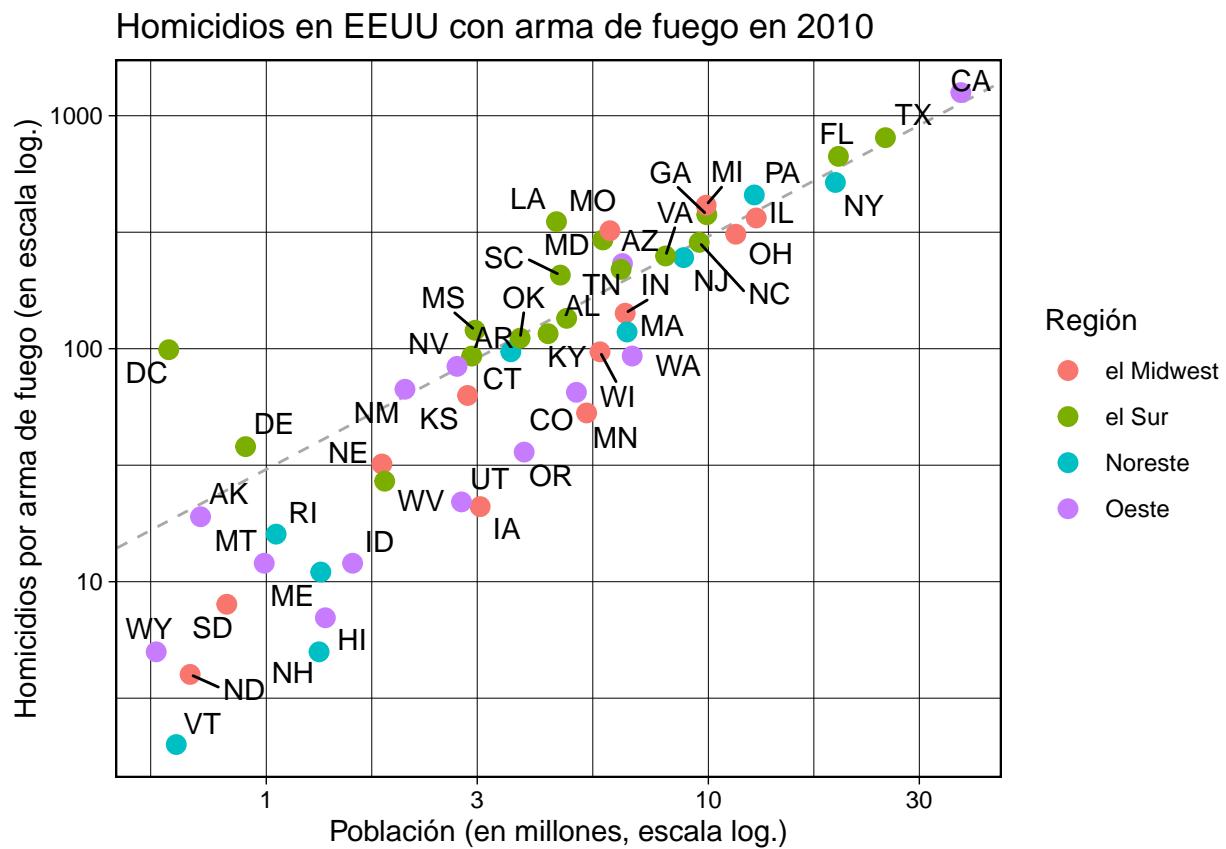
library(dslabs)
t <- murders |>
  summarise(tasa = sum(total) / sum(population) * 10^6) |>
  pull(tasa)

murders |>
  mutate(region=case_when(
    region == "Northeast" ~ "Noreste",
    region == "North Central" ~ "el Midwest",
    region == "West" ~ "Oeste",
    region == "South" ~ "el Sur"))%>%
  ggplot(aes(population/10^6, total)) +
  geom_abline(intercept = log10(t), lty = 2, color = "darkgrey") +
  geom_point(aes(col = region), size = 3) +
```

```

geom_text_repel(aes(label = abb)) +
scale_x_log10() +
scale_y_log10() +
labs(title = "Homicidios en EEUU con arma de fuego en 2010",
x = "Población (en millones, escala log.)",
y = "Homicidios por arma de fuego (en escala log.)",
color = "Región") +
theme_linedraw()

```



He aquí los pasos entonces que necesitábamos para recrear la imagen arriba.

Datos del Covid

Durante la pandemia del Coronavirus de 2019, todos pasamos por bastantes cosas, entre ellas, tratar de entender el fenómeno inaudito en nuestras vidas, que era una mortífera y peligrosa pandemia, de la cual desconocíamos en general muchas características. Esto llevó a los gobiernos del mundo a tomar distintos tipos de acciones y nos informábamos con gráficas así como números y tablas del progreso de la enfermedad y su avance a través de los países del planeta, así como la variable tasa de mortalidad que la acompañaba. Varios científicos lanzaron programas para recoger y expresar estos datos al público general para mantenernos todos bien informados. En mi caso tomé código y datos que se recogían a menudo y produje por unos meses varios tipos de gráficos para representar el desarrollo de la pandemia. Aquí doy una versión del código que utilicé para estos fines (modificando las fechas ya que se siguió recopilando esa información mucho después de cuando dejara de actualizar esos gráficos).

```

#devtools::install_github("RamiKrispin/coronavirus", force = TRUE)
library(coronavirus)
library(scales)

## 
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':
## 
##     discard

## The following object is masked from 'package:readr':
## 
##     col_factor

coronavirus<-coronavirus
the_caption = "Fuente: OMS y varios via la Universidad John Hopkins y el paquete de R 'coronavirus' de"
top_countries <- coronavirus %>%
  filter(date <= as.Date("2020-07-31")) %>%
  filter(type == "confirmed") %>%
  group_by(country) %>%
  summarise(cases = sum(cases)) %>%
  top_n(10, wt = cases)

d2 <- coronavirus %>%
  filter(date <= as.Date("2020-07-31")) %>% # Filtro para datos dentro de 2020
  group_by(date, country, type) %>%
  summarise(cases = sum(cases)) %>%
  group_by(date, country) %>%
  spread(type, cases) %>%
  arrange(date) %>%
  group_by(country) %>%
  mutate(cfr_cumulative = cumsum(death) / cumsum(confirmed)) %>%
  filter(!is.na(cfr_cumulative)) %>%
  ungroup() %>%
  inner_join(top_countries, by = "country")

## `summarise()` has grouped output by 'date', 'country'. You can override using
## the ` `.groups` argument.

summary(as.factor(d2$country))

```

	Brazil	Chile	India	Iran	Mexico
##	157	160	184	164	155
##	Peru	Russia	South Africa	United Kingdom	US
##	148	183	149	184	192

```

today<-as.Date("2020-08-06")
x_limits <- c(today, NA)

```

```
top_countries
```

```

## # A tibble: 10 x 2
##   country      cases
##   <chr>        <dbl>
## 1 Brazil      2670451
## 2 Chile       355667
## 3 India       1695988
## 4 Iran        304204
## 5 Mexico      424637
## 6 Peru         407492
## 7 Russia      838461
## 8 South Africa 493183
## 9 US          4548497
## 10 United Kingdom 304789

```

Vemos que los países están en inglés pero queremos que aparezcan en nuestro gráfico en español. Podemos modificar la información con tidyverse:

```

d2<- d2 %>%
  mutate(country=recode(country, US = "EEUU", Russia ="Rusia", Mexico = "México", Brazil = "Brasil",
                        "United Kingdom" = "Reino Unido", Japan="Japón",
                        Italy = "Italia", Iran = "Irán", Peru = "Perú",
                        "South Africa"= "Sudáfrica"))
d2<- d2 %>%
  mutate(cfr_cumulativeperc=round(cfr_cumulative*100,2))

```

Finalmente, expresamos el gráfico buscado. Paso los datos de mapeo estético como opciones globales y le añado las capas: líneas para series de tiempo, etiquetas automáticamente posicionadas (pero editadas para especificar dónde las quiero y añadir información adicional), transformaciones numéricas, información adicional de selección colores, una extensión del marco para acomodar las etiquetas, así como los títulos que deseara utilizar:

```

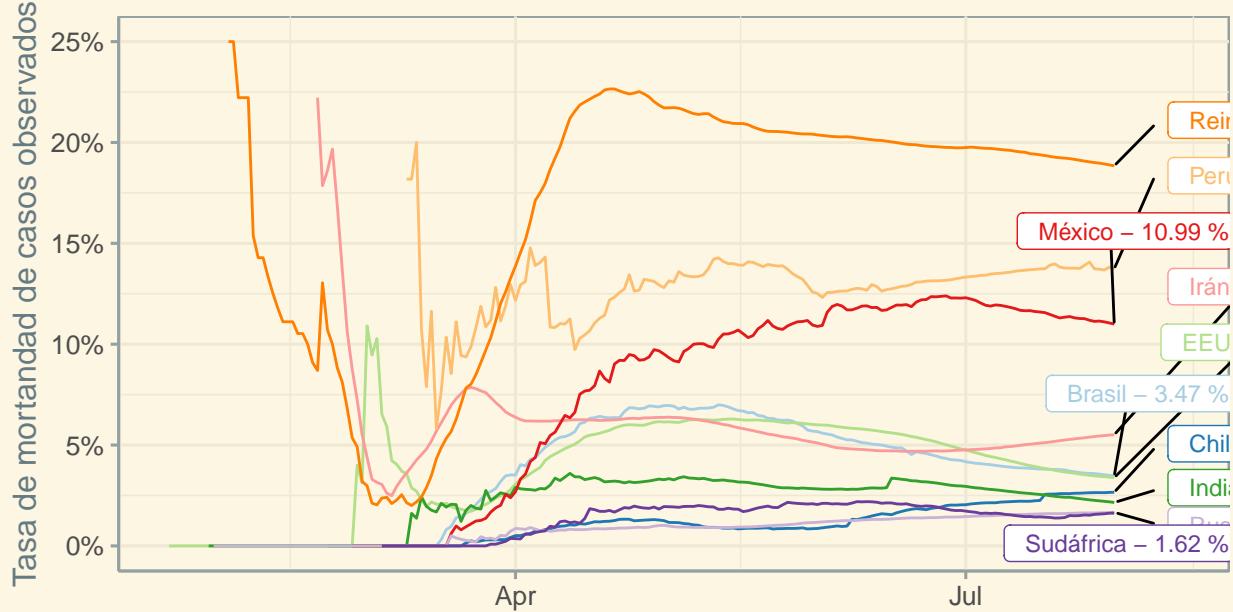
d2 %>%
  ggplot(aes(x = date, y = cfr_cumulative, colour = country)) +
  geom_line() +
  geom_label_repel(data = filter(d2, date == max(date)), aes(label = paste("",country,"-",cfr_cumulative,
                                                          hjust = 1, size = 3, xlim = x_limits, segment.color="black" ) +
    scale_y_continuous(label = percent_format(accuracy = 1), limits = c(0, .25)) +
    scale_colour_brewer(type = 'qual', palette = 'Paired', direction = 1) +
    expand_limits(x = max(d2$date) + 13) +
    labs(caption = the_caption,
         x = "",
         y = "Tasa de mortalidad de casos observados",
         title = "Fatalidad de casos de COVID-19 en los diez países con más casos diagnosticados",
         subtitle = "Los diez países con mayor cantidad de casos diagnosticados. El caso de Irán ha sido "
    Un alto grado de incertidumbre refleja denominadores que cambian con fluidez, y el desconocimiento de ca
    theme_solarized()+
    theme(legend.position = "none"))

## Warning: Removed 6 rows containing missing values or values outside the scale range
## (`geom_line()`).

```

Fatalidad de casos de COVID-19 en los diez países con más casos

Los diez países con mayor cantidad de casos diagnosticados. El caso de Irán ha sido muy notorio. Un alto grado de incertidumbre refleja denominadores que cambian con fluidez,



Fuente: OMS y varios via la Universidad John Hopkins y el paquete de R 'coronavirus' de Rami Krispin. Análisis de Rashid C.J. Marcano Rivera, modificando código provisto por <http://freerangestats.info> (Peter Ellis)

Datos del censo

En este bloque introductorio, utilizamos los datos del Censo de Estados Unidos para obtener variables demográficas y socioeconómicas de Puerto Rico. Los datos del Censo nos permiten comprender mejor las características de la población, como ingresos, educación, vivienda, y más, a lo largo del tiempo. Para este ejercicio, hacemos uso de `tidycensus`, una poderosa herramienta que facilita la descarga y manipulación de los datos censales dentro del entorno de R.

El primer paso es cargar las bibliotecas necesarias, como `tidyverse`, `tidycensus` y `sf`. La combinación de estas bibliotecas nos permitirá no solo acceder a los datos, sino también analizarlos y visualizarlos espacialmente en mapas. Es importante también tener configurada una clave API de la Oficina del Censo para poder acceder a los datos.

Luego, se configura el entorno con algunas opciones útiles:

- `scipen = 999`: Evita el uso de notación científica en los números, lo cual facilita la lectura de resultados.
- `tigris_class = "sf"`: Permite manejar las geometrías de manera eficiente mediante `sf`, una clase para manejar datos geoespaciales. A continuación, se usan las funciones de `tidycensus` para cargar variables específicas de diversos conjuntos de datos censales, como la Encuesta de la Comunidad Americana (ACS), la Encuesta de la Comunidad de Puerto Rico (PRCS, que se obtiene aquí vía la función de ACS) y los datos del Censo Decenal. Cada conjunto de datos proporciona información clave sobre diferentes períodos: anual, quinquenal y decenal. Finalmente, mostramos las variables disponibles que podemos utilizar para análisis posteriores.

```

library(tidyverse)
library(tidycensus)

## Warning: package 'tidycensus' was built under R version 4.3.3

library(sf)

## Warning: package 'sf' was built under R version 4.3.3

## Linking to GEOS 3.11.0, GDAL 3.5.3, PROJ 9.1.0; sf_use_s2() is TRUE

#census_api_key(INSERTE SU LLAVE DEL CENSO AQUÍ)
options(scipen=999)
options(tigris_class = "sf")
#https://api.census.gov/data.html
census_variables <- load_variables(year = 2020, dataset = "acs5", cache = TRUE)
#ensus_variablesf1 <- load_variables(year = 2020, dataset = "sf1", cache = TRUE)
#census_variablesf2 <- load_variables(year = 2020, dataset = "sf2", cache = TRUE)
#census_variables4Redist <- load_variables(year = 2020, dataset = "pl", cache = TRUE)
#v00 <- load_variables(2000, "sf3", cache = TRUE)
v18 <- load_variables(2018, "acs5", cache = TRUE)
v10 <- load_variables(2010, "sf1", cache = TRUE)
census_variables

## # A tibble: 27,850 x 4
##   name      label          concept      geography
##   <chr>     <chr>        <chr>        <chr>
## 1 B01001A_001 Estimate!!Total:    SEX BY AGE (W~ tract
## 2 B01001A_002 Estimate!!Total:!!Male:   SEX BY AGE (W~ tract
## 3 B01001A_003 Estimate!!Total:!!Male:!!Under 5 years SEX BY AGE (W~ tract
## 4 B01001A_004 Estimate!!Total:!!Male:!!5 to 9 years SEX BY AGE (W~ tract
## 5 B01001A_005 Estimate!!Total:!!Male:!!10 to 14 years SEX BY AGE (W~ tract
## 6 B01001A_006 Estimate!!Total:!!Male:!!15 to 17 years SEX BY AGE (W~ tract
## 7 B01001A_007 Estimate!!Total:!!Male:!!18 and 19 years SEX BY AGE (W~ tract
## 8 B01001A_008 Estimate!!Total:!!Male:!!20 to 24 years SEX BY AGE (W~ tract
## 9 B01001A_009 Estimate!!Total:!!Male:!!25 to 29 years SEX BY AGE (W~ tract
## 10 B01001A_010 Estimate!!Total:!!Male:!!30 to 34 years SEX BY AGE (W~ tract
## # i 27,840 more rows

v10

## # A tibble: 8,959 x 3
##   name      label          concept
##   <chr>     <chr>
## 1 H001001 Total:           HOUSING UNITS
## 2 H002001 Total:           URBAN AND RURAL
## 3 H002002 Total!!Urban:    URBAN AND RURAL
## 4 H002003 Total!!Urban!!Inside urbanized areas URBAN AND RURAL
## 5 H002004 Total!!Urban!!Inside urban clusters URBAN AND RURAL
## 6 H002005 Total!!Rural:    URBAN AND RURAL

```

```

## 7 H002006 Total!!Not defined for this file      URBAN AND RURAL
## 8 H003001 Total                                OCCUPANCY STATUS
## 9 H003002 Total!!Occupied                     OCCUPANCY STATUS
## 10 H003003 Total!!Vacant                      OCCUPANCY STATUS
## # i 8,949 more rows

```

v18

```

## # A tibble: 26,997 x 4
##   name      label            concept      geography
##   <chr>     <chr>          <chr>        <chr>
## 1 B00001_001 Estimate!!Total    UNWEIGHTED SAMP~ block gr~
## 2 B00002_001 Estimate!!Total    UNWEIGHTED SAMP~ block gr~
## 3 B01001A_001 Estimate!!Total   SEX BY AGE (WHI~ tract
## 4 B01001A_002 Estimate!!Total!!Male  SEX BY AGE (WHI~ tract
## 5 B01001A_003 Estimate!!Total!!Male!!Under 5 years  SEX BY AGE (WHI~ tract
## 6 B01001A_004 Estimate!!Total!!Male!!5 to 9 years  SEX BY AGE (WHI~ tract
## 7 B01001A_005 Estimate!!Total!!Male!!10 to 14 years  SEX BY AGE (WHI~ tract
## 8 B01001A_006 Estimate!!Total!!Male!!15 to 17 years  SEX BY AGE (WHI~ tract
## 9 B01001A_007 Estimate!!Total!!Male!!18 and 19 years  SEX BY AGE (WHI~ tract
## 10 B01001A_008 Estimate!!Total!!Male!!20 to 24 years  SEX BY AGE (WHI~ tract
## # i 26,987 more rows

```

Podemos pasar por las variables en catálogo y seleccionar la que nos llame más la atención. En este caso, quiero revisar para algún futuro trabajo de investigación datos sobre el ingreso medio de los distritos representativos de Puerto Rico para los años 2018 y 2019 (que en el Censo aparecen como geografía tipo **state legislative district (lower chamber)**). Usaré la función `get_acs`, podemos ver qué hace:

?get_acs

Para efectos estadísticos Puerto Rico entra como estado en estos datos. Uso entonces la variable `B19013_001` que da los ingresos medianos por hogar.

```

prmedian_2019 <- get_acs(geography = "state legislative district (lower chamber)",
                           variables = c(medincome = "B19013_001"),
                           state = "PR",
                           year = 2019)

```

Getting data from the 2015–2019 5-year ACS

```

prmedian_2018 <- get_acs(geography = "state legislative district (lower chamber)",
                           variables = c(medincome = "B19013_001"),
                           state = "PR",
                           year = 2018)

```

Getting data from the 2014–2018 5-year ACS

```
head(prmedian_2018)
```

A tibble: 6 x 5

```

##   GEOID NAME                                variable estimate    moe
##   <chr> <chr>                                <chr>      <dbl> <dbl>
## 1 72001 State House District 1 (2018), Puerto Rico medincome     18336    955
## 2 72002 State House District 2 (2018), Puerto Rico medincome     18343    884
## 3 72003 State House District 3 (2018), Puerto Rico medincome    20315   1394
## 4 72004 State House District 4 (2018), Puerto Rico medincome    33471   1586
## 5 72005 State House District 5 (2018), Puerto Rico medincome    23274   1619
## 6 72006 State House District 6 (2018), Puerto Rico medincome    32921   1581

```

```

pr_median_combined <- prmedian_2019 %>%
  mutate(year = 2019) %>%
  bind_rows(
    prmedian_2018 %>% mutate(year = 2018)
  )
head(pr_median_combined)

```

```

## # A tibble: 6 x 6
##   GEOID NAME                                variable estimate    moe year
##   <chr> <chr>                                <chr>      <dbl> <dbl> <dbl>
## 1 72001 State House District 1 (2018), Puerto Rico medinco~     18523  1142 2019
## 2 72002 State House District 2 (2018), Puerto Rico medinco~     19004   995 2019
## 3 72003 State House District 3 (2018), Puerto Rico medinco~     20655  1400 2019
## 4 72004 State House District 4 (2018), Puerto Rico medinco~     33678  1519 2019
## 5 72005 State House District 5 (2018), Puerto Rico medinco~     25912  1733 2019
## 6 72006 State House District 6 (2018), Puerto Rico medinco~     33977  1491 2019

```

Habiendo obtenido las tablas, digamos que no quiero hacer sólo un análisis numérico a ojo. El ojo y mente humana rara vez puede determinar mucho al ver una colección grande de estos datos rápidamente. Pero al visualizar, podemos mejorar algo la velocidad en lo que entendemos lo que está ante nos.

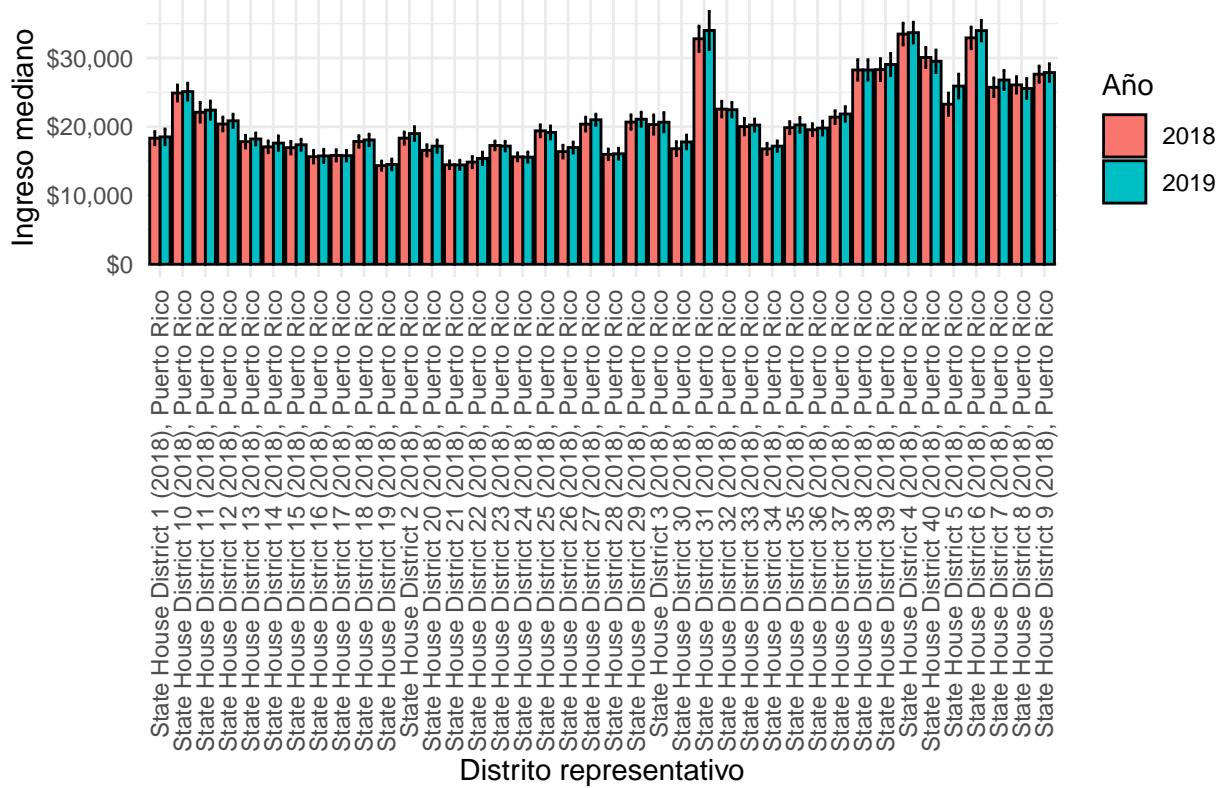
```

pr_median_combined <- pr_median_combined %>%
  filter(!is.na(estimate)) # Filter out rows where the estimate is NA

pr_median_combined |>
  ggplot(aes(x = NAME, y = estimate, fill = factor(year))) +
  geom_bar(stat = "identity", position = position_dodge(), color = "black") +
  geom_errorbar(aes(ymin = estimate - moe, ymax = estimate + moe),
                position = position_dodge(0.9), width = 0.25) +
  scale_y_continuous(labels = scales::dollar_format()) +
  labs(
    title = "Ingreso mediano por distrito legislativo en Puerto Rico (2018 vs 2019)",
    x = "Distrito representativo",
    y = "Ingreso mediano",
    fill = "Año"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1) # Rotar nombres de distritos 90 grados
  )

```

Ingreso mediano por distrito legislativo en Puerto Rico (2018 vs 2019)



Finalmente, usaremos esta información para generar un mapa geospatial con ggplot. En este caso tomamos como parte de los pasos una transformación de coordenadas para que el mapeo entienda dónde posicionar las geometrías en el mapeo estético.

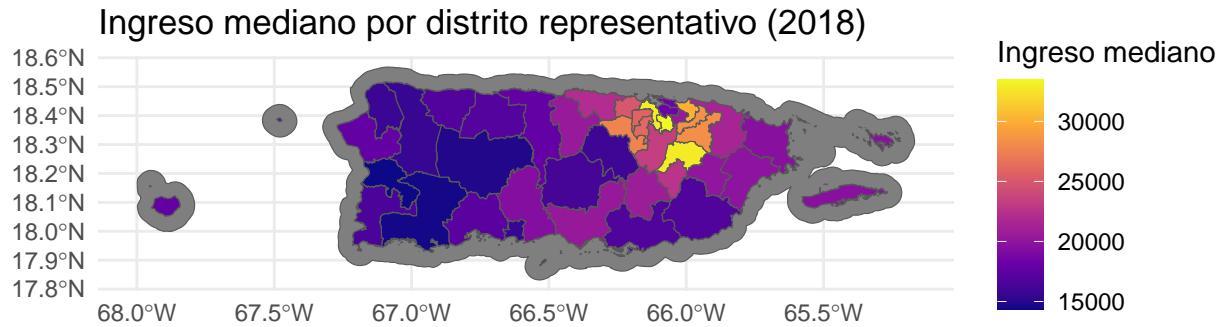
```
library(tigris)

## To enable caching of data, set `options(tigris_use_cache = TRUE)`
## in your R script or .Rprofile.

pr_legislative_districts <- state_legislative_districts(state = "PR", house = "lower", year = 2018)

## | 

pr_legislative_districts <- st_transform(pr_legislative_districts, crs = "EPSG:4326") # ajustar proyección
pr_median_2018_geo <- pr_legislative_districts %>%
  left_join(prmedian_2018, by = c("GEOID" = "GEOID"))
pr_median_2018_geo |>
  ggplot() +
  geom_sf(aes(fill = estimate)) +
  scale_fill_viridis_c(option = "plasma", name = "Ingreso mediano") +
  theme_minimal() +
  labs(
    title = "Ingreso mediano por distrito representativo (2018)",
    fill = "Ingreso mediano"
  )
```



En este bloque próximo, utilizamos datos espaciales y demográficos para mapear la población de Puerto Rico a nivel de trácticos censales del censo de 2020. Combinamos dos fuentes de información: una capa espacial con los límites geográficos de los trácticos (la forma de los polígonos en un *shapefile*: <https://electionspuertorico.org/datos/2020/#MAPA>) y los datos poblacionales del Censo Decenal de 2020.

1. El shapefile nos proporciona los límites de los trácticos que necesitamos para asignar la información demográfica correctamente.
 - **st_read()**: Carga el shapefile, aplicando la codificación `latin1` para asegurarnos de que los caracteres especiales (acentos, diéresis, eñes) en los nombres se lean correctamente.
 - **st_set_crs() y st_transform()**: Aquí ajustamos el CRS (sistema de referencia de coordenadas) de la capa geográfica. Primero asignamos el CRS EPSG:4326 (coordenadas geográficas), y luego transformamos la capa a EPSG:3920 (un CRS proyectado que es común para Puerto Rico) para asegurar que todas las capas estén en el mismo sistema de referencia.
2. **Datos de población:** Usamos `tidycensus` para obtener los datos de población de Puerto Rico a nivel de trácticos censales del censo de 2020.
 - **get_decennial()**: Esta función descarga los datos de población (variable `P1_001N`, que representa el total de la población). La opción `geometry = TRUE` incluye la geometría de los trácticos, lo cual nos permite hacer un análisis espacial directamente.
 - **st_transform()**: Aplicamos la misma transformación de CRS a los datos de población para que coincidan con la capa geográfica de los trácticos.
3. **Unión espacial:** Utilizamos `st_join()` para combinar los datos espaciales y de población, uniendo la capa de trácticos con los datos del censo según el campo `GEOID`. Esto nos permite representar los datos poblacionales en el mapa.

```

# Este archivo lo descargué el 9 de agosto de 2022 desde https://electionspuertorico.org/datos/2020/PR2020TV.shp
# folder <- "~/Library/CloudStorage/OneDrive-IndianaUniversity/Elecciones/Rashid and Brevin/"

pr_unidad_shp_filename <- paste(folder, "PR2020Shp/PR2020TV.shp", sep= "")
pr_unidades <- st_read(pr_unidad_shp_filename, options = "ENCODING=latin1")

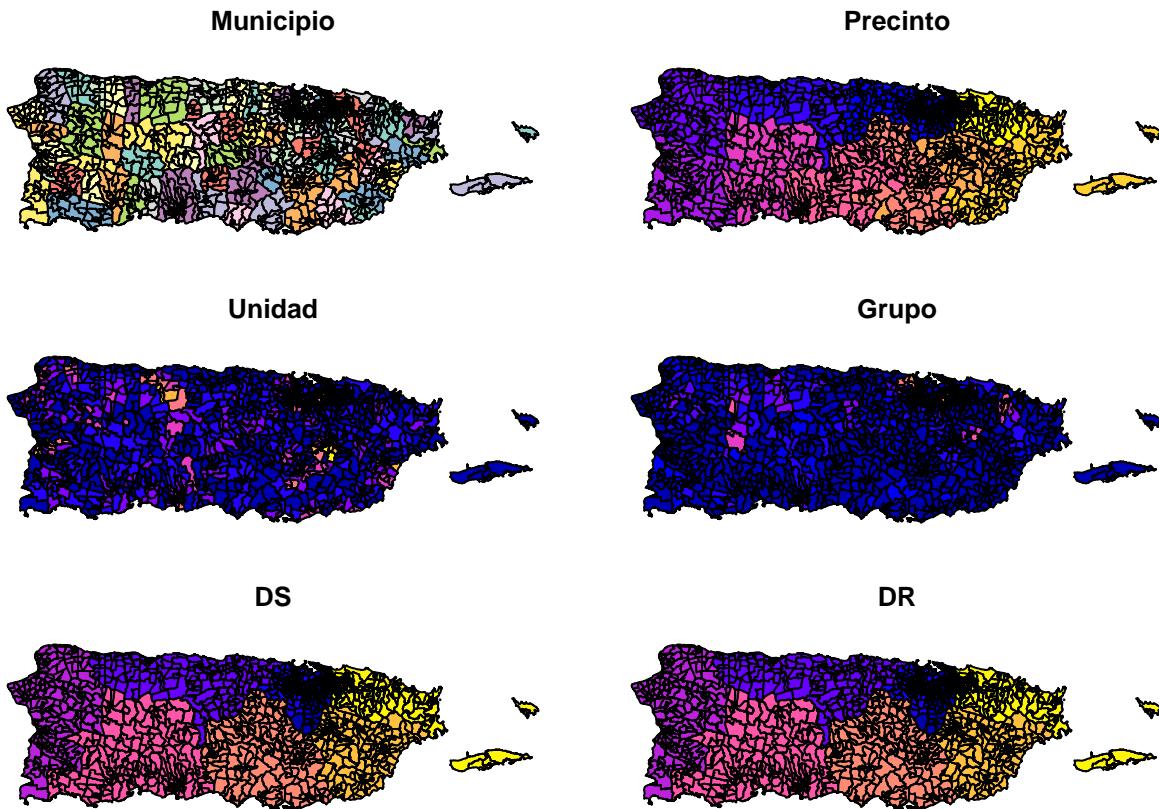
## options:      ENCODING=latin1
## Reading layer `PR2020TV' from data source
##   `/Users/rashid/Library/CloudStorage/OneDrive-IndianaUniversity/Elecciones/Rashid and Brevin/PR2020TV.shp'
##   using driver `ESRI Shapefile'
## Simple feature collection with 1365 features and 6 fields
## Geometry type: MULTIPOLYGON
## Dimension: XY
## Bounding box: xmin: -67.2713 ymin: 17.92684 xmax: -65.2442 ymax: 18.51609
## Geodetic CRS: WGS 84

st_set_crs(pr_unidades, "EPSG:4326")

## Simple feature collection with 1365 features and 6 fields
## Geometry type: MULTIPOLYGON
## Dimension: XY
## Bounding box: xmin: -67.2713 ymin: 17.92684 xmax: -65.2442 ymax: 18.51609
## Geodetic CRS: WGS 84
## First 10 features:
##   Municipio Precinto Unidad Grupo DS DR           geometry
## 1 Barranquitas     70      5    3 6 26 MULTIPOLYGON (((-66.30361 1...
## 2 Cabo Rojo        46      2    3 4 20 MULTIPOLYGON (((-67.14746 1...
## 3 Cabo Rojo        46      3    3 4 20 MULTIPOLYGON (((-67.16632 1...
## 4 Cabo Rojo        46      9    2 4 20 MULTIPOLYGON (((-67.1375 18...
## 5 Cabo Rojo        46      6    3 4 20 MULTIPOLYGON (((-67.16957 1...
## 6 San Germán       44      3    1 4 20 MULTIPOLYGON (((-67.08108 1...
## 7 San Germán       43      7    2 4 19 MULTIPOLYGON (((-67.02885 1...
## 8 San Germán       43      6    2 4 19 MULTIPOLYGON (((-67.04865 1...
## 9 San Germán       44      4    1 4 20 MULTIPOLYGON (((-67.06009 1...
## 10 San Germán      43      2    1 4 19 MULTIPOLYGON (((-67.08731 1...

pr_unidades <- pr_unidades %>% st_transform(crs = "EPSG:3920")
plot(pr_unidades)

```



```
pr_population <- get_decennial(
  geography = "tract",
  variables = "P1_001N",
  state = "72",    # FIPS code for Puerto Rico
  year = 2020,
  geometry = TRUE
)
```

```
## Getting data from the 2020 decennial Census
```

```
## Downloading feature geometry from the Census website. To cache shapefiles for use in future sessions
```

```
## Using the PL 94-171 Redistricting Data Summary File
```

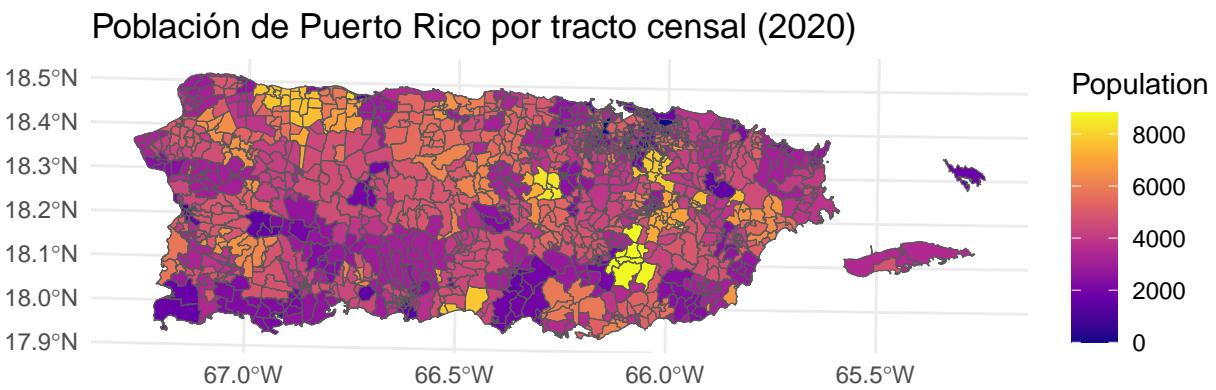
```
## |
```

```
## Note: 2020 decennial Census data use differential privacy, a technique that
## introduces errors into data to preserve respondent confidentiality.
## i Small counts should be interpreted with caution.
## i See https://www.census.gov/library/fact-sheets/2021/protecting-the-confidentiality-of-the-2020-cen
## This message is displayed once per session.
```

```

pr_population <- st_transform(pr_population, crs = "EPSG:3920")
pr_unidades <- pr_unidades %>%
  st_join(pr_population, by = c("GEOID" = "GEOID")) # Assuming GEOID is the common field
pr_unidades|>ggplot() +
  geom_sf(aes(fill = value)) + # `value` is the population count
  scale_fill_viridis_c(option = "plasma", name = "Population") +
  theme_minimal() +
  labs(title = "Población de Puerto Rico por tracto censal (2020)")

```



En este bloque, generamos un mapa que muestra la **población por tracto censal** en Puerto Rico. Cada trácto está coloreado según su población, lo que nos permite visualizar cómo se distribuye la población a lo largo de la isla.

- **geom_sf(aes(fill = value)):** Aquí, visualizamos los tráctos censales y asignamos un color a cada uno de acuerdo con su población (**value**).
- **scale_fill_viridis_c():** Utilizamos una escala de color continua, con la opción **plasma**, para mejorar la legibilidad de los datos.
- **labs():** Añadimos un título al gráfico y una leyenda clara para que el lector entienda fácilmente la visualización.

Recapitulando

Qué aprendimos en este taller inicial

Hoy aprendimos varias cosas en R:

- Aprendimos principios de visualización
- Lenguaje de gramática de gráficos
- Uso de capas para añadir complejidad visual a los datos
- Gráficas con datos complejos
- Uso de `tidycensus` para descargar datos censales
- Gráficos con mapas

Continuamos el próximo viernes

En los talleres que vienen continuaremos ahondando en operaciones estadísticas, así como estadística inferencial, y finalmente gráficos que ayuden a entender y diagnosticar los modelos estadísticos que usaremos. Gracias por asistir hoy.