Clase_ggplot_2

Joshua Kock

3/29/2019

```
library(tidyverse)
```

```
## -- Attaching packages ----------------------------------------------------------------- tidyverse 1

## v ggplot2 3.1.0       v purrr   0.3.2
## v tibble  2.1.1       v dplyr   0.8.0.1
## v tidyr   0.8.3       v stringr 1.4.0
## v readr   1.3.1       v forcats 0.4.0

## -- Conflicts -------------------------------------------------------------------- tidyverse_conflic
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(gapminder)

theme_set(theme_light())
```

Introduccion

Continuaremos desarrollando un flujo de trabajo relacionado con ggplot y expandiendo en la posibilidad de herramientas y funciones que puedes hacer con ella.

El código casi nunca funciona correctamente la primera vez que lo escribes. Esta es la razón principal por la que, al aprender un nuevo lenguaje, es importante escribir los ejercicios y seguirlos manualmente. Te da una mejor idea de cómo funciona la sintaxis del lenguaje. Si algo salió mal, puedes averiguar por qué sucedió.

Los errores pueden ser con:

- estructura interna de tus datos (grouping)
- como dividir tus datos en pedazos para graficar (faceting)
- como tranformar datos y calculos para producir tu grafico (tranforming)

Hay que tener en cuenta que usualmente la primera grafica nunca sera lo que tienes en mente, deberas hacer varias iteraciones de la grafica hasta obtener lo que quieres comunicar visualment.

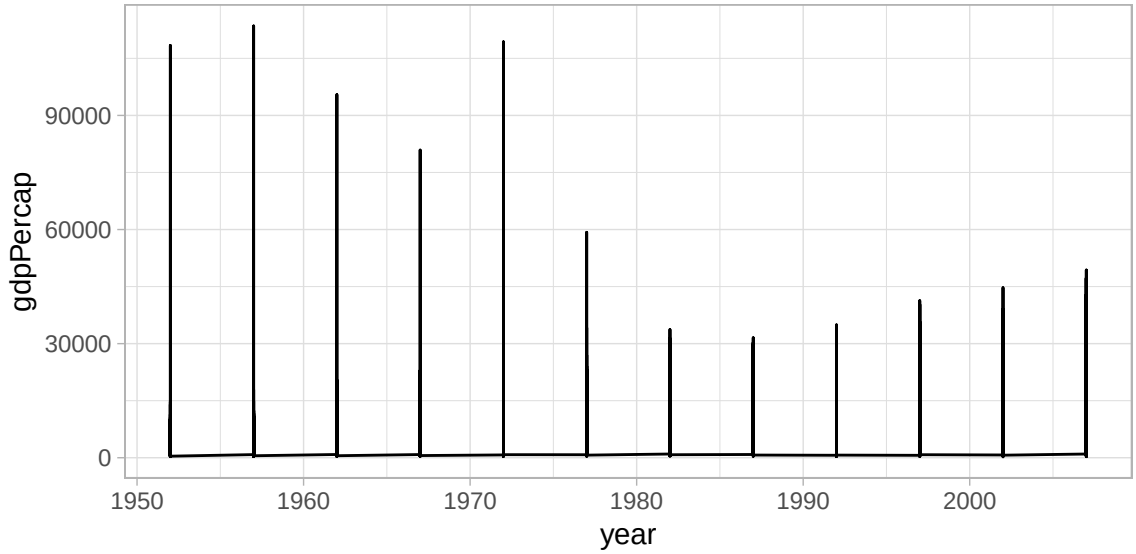
Datos agrupados estetica de "group".

Vamos a continuar con `gapminder`. Imaginen que queremos hacer una grafica de la tendencia de la PIB per capita de cada pais. Tenemos que mapear `year` al eje-x, `gdpPercap` en eje-y.

```
head(gapminder)
```

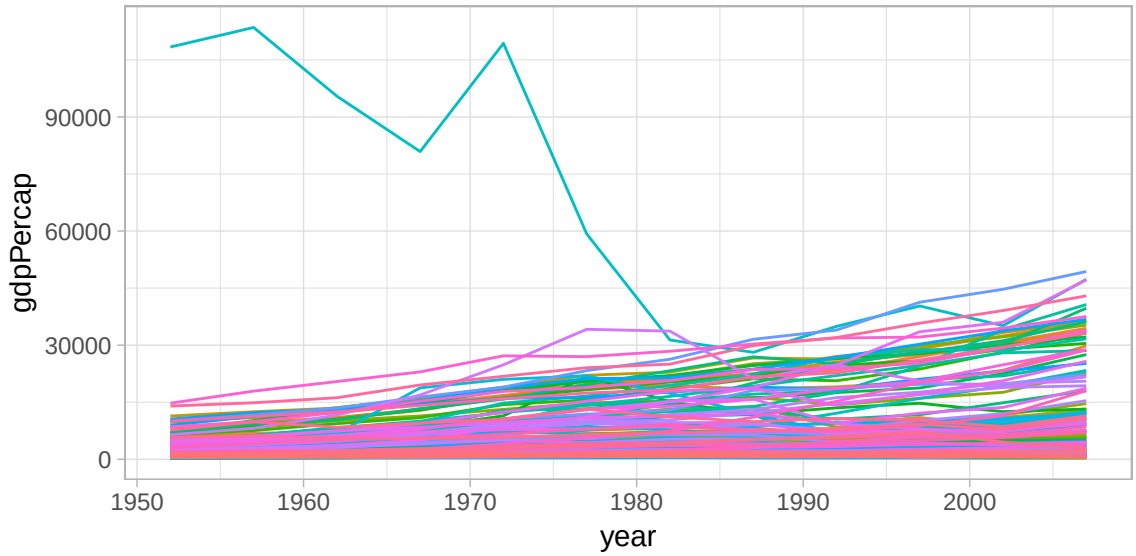
```
## # A tibble: 6 x 6
##   country     continent  year lifeExp      pop gdpPercap
##   <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
## 1 Afghanistan Asia       1952    28.8  8425333      779.
## 2 Afghanistan Asia       1957    30.3  9240934      821.
## 3 Afghanistan Asia       1962    32.0 10267083      853.
## 4 Afghanistan Asia       1967    34.0 11537966      836.
## 5 Afghanistan Asia       1972    36.1 13079460      740.
## 6 Afghanistan Asia       1977    38.4 14880372      786.
```

```
ggplot(data = gapminder, aes(x = year, y = gdpPercap)) +
  geom_line()
```



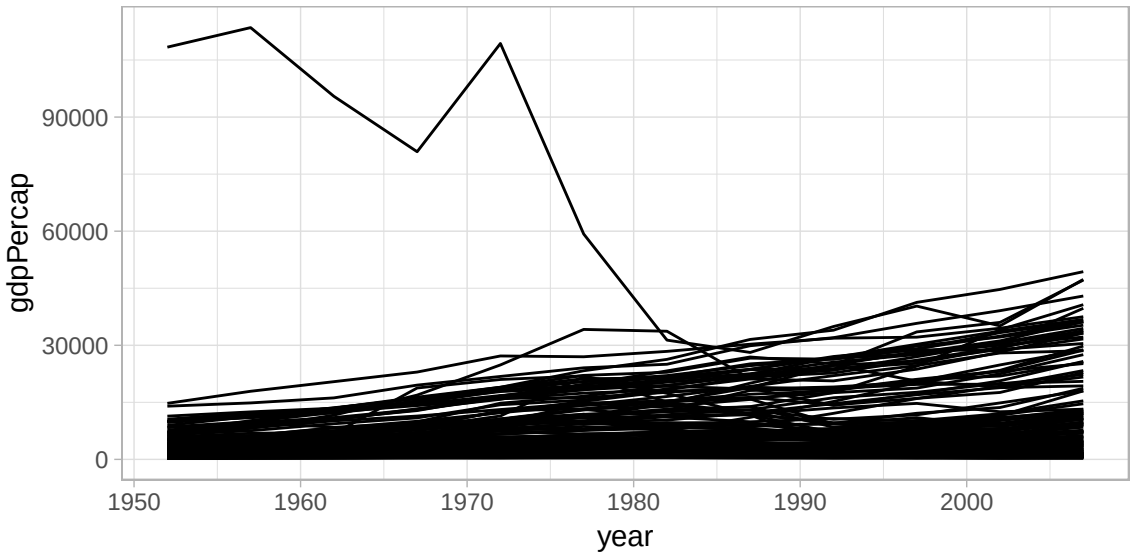
Que podemos hacer para arreglarlo

```
ggplot(data = gapminder, aes(x = year, y = gdpPercap, colour = country)) +
  geom_line(show.legend = FALSE)
```



La forma correcta es usando la estetica de grupo, te dara una estructura por pais especificamente

```
ggplot(data = gapminder, aes(x = year, y = gdpPercap)) +
  geom_line(aes(group = country))
```



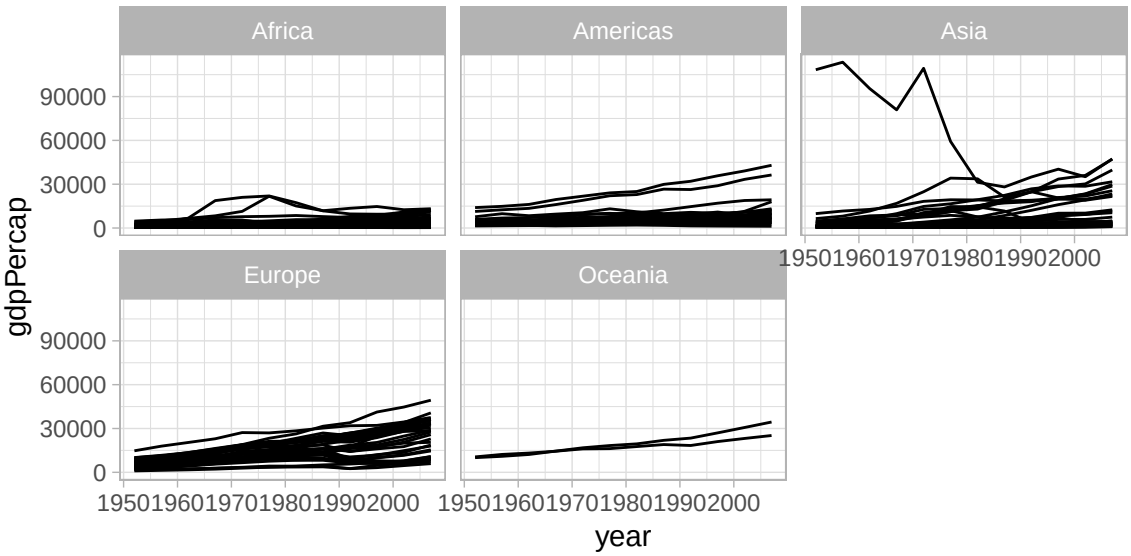
La estética del grupo generalmente solo es necesaria cuando la información de agrupación que necesitas para informar a ggplot no está integrada en las variables que se asignan. Por ejemplo, cuando graficando por continente, con mapear el color al continente era suficiente para obtener la respuesta correcta, porque el continente ya es una variable categórica con pocas categorias, con una agrupación es clara.

La Faceta "Facet" para hacer multiples pequeños.

La grafica anterior tiene muchas lineas y no podemos distinguir entre los paises, no seria ideal para comunicar la tendencia de PIC per capita de los paises. Una opcion es pasar una faceta a una tercera variable, creando algo que se llama multiples pequeños. Se grafica un panel diferente para cada valor de la variables en la faceta.

La faceta no es un geom/geometria, la faceta es una forma de organizar tu geometria. Para esto utilizaremos la funcion: `facet_wrap()`

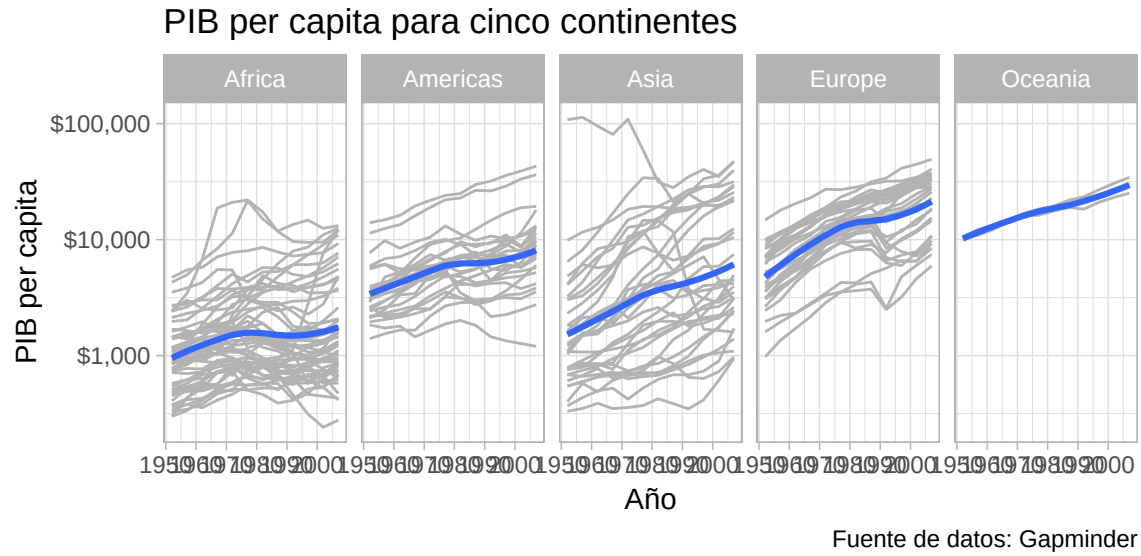
```
ggplot(data = gapminder, aes(x = year, y = gdpPercap)) +
  geom_line(aes(group = country)) +
  facet_wrap(~ continent)
```



La faceta minimiza la duplicacion de ejes, cada faceta esta etiquetada arriba.

Vamos hacer una grafica mas elegante mostrando la tendencia por continente.

```
ggplot(data = gapminder, mapping = aes(x = year, y = gdpPercap)) +
  geom_line(color = "gray70", aes(group = country)) +
  geom_smooth(size = 1.1, method = "loess", se = FALSE) +
  scale_y_log10(labels = scales::dollar) +
  facet_wrap(~ continent, ncol = 5) +
  labs(x = "Año",
       y = "PIB per capita",
       title = "PIB per capita para cinco continentes",
       caption = "Fuente de datos: Gapminder")
```



Vamos a trabajar con estos datos que provienen de una encuenta de social realizada en el 2016 para mas informacion de las variables puedes consultar: http://gss.norc.org/Get-Documentation

```
base_gss_sm <- read_csv("https://raw.githubusercontent.com/vizual-wanderer/6071402_Electiva_II/master/Ba
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   year = col_double(),
##   id = col_double(),
##   ballot = col_double(),
##   age = col_double(),
##   childs = col_double(),
##   sibs = col_double(),
##   pres12 = col_double(),
##   wtssall = col_double(),
##   obama = col_double()
## )
```

```
## See spec(...) for full column specifications.
```

```
base_gss_sm
```

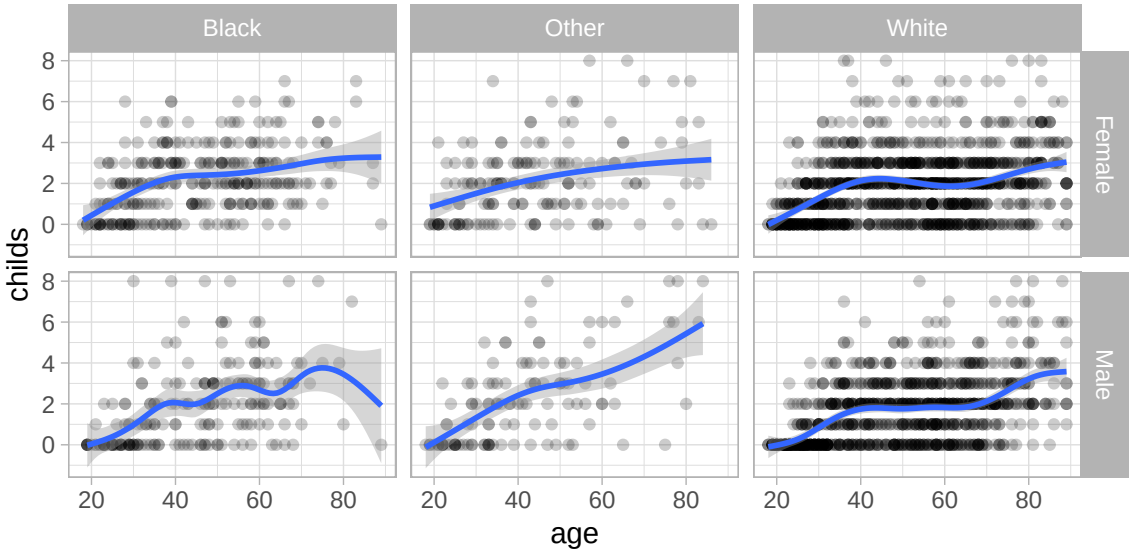
```
## # A tibble: 2,867 x 32
##     year    id ballot   age childs  sibs degree race  sex   region income16
##    <dbl> <dbl>  <dbl> <dbl>  <dbl> <dbl> <chr>  <chr> <chr> <chr>  <chr>
##  1  2016     1      1    47      3     2 Bache~ White Male  New E~ $170000~
##  2  2016     2      2    61      0     3 High ~ White Male  New E~ $50000 ~
##  3  2016     3      3    72      2     3 Bache~ White Male  New E~ $75000 ~
##  4  2016     4      1    43      4     3 High ~ White Fema~ New E~ $170000~
##  5  2016     5      3    55      2     2 Gradu~ White Fema~ New E~ $170000~
##  6  2016     6      2    53      2     2 Junio~ White Fema~ New E~ $60000 ~
##  7  2016     7      1    50      2     2 High ~ White Male  New E~ $170000~
##  8  2016     8      3    23      3     6 High ~ Other Fema~ Middl~ $30000 ~
##  9  2016     9      1    45      3     5 High ~ Black Male  Middl~ $60000 ~
## 10  2016    10      3    71      4     1 Junio~ White Male  Middl~ $60000 ~
## # ... with 2,857 more rows, and 21 more variables: relig <chr>,
## #   marital <chr>, padeg <chr>, madeg <chr>, partyid <chr>,
## #   polviews <chr>, happy <chr>, partners <chr>, grass <chr>,
## #   zodiac <chr>, pres12 <dbl>, wtssall <dbl>, income_rc <chr>,
## #   agegrp <chr>, ageq <chr>, siblings <chr>, kids <chr>, religion <chr>,
## #   bigregion <chr>, partners_rc <chr>, obama <dbl>
```

Ejercicio:

Realicen un grafico de numero hijos que tiene cada sujeto por la edad del sujeto: `childs` y `age`, estas hacer una faceta por sexo y raza `sex` y `race` verificando la tendencia para cada grupo. Recomendacion de utilizar `facet_grid`.

```
ggplot(data = base_gss_sm,
       mapping = aes(x = age, y = childs)) +
  geom_point(alpha = 0.2) +
  geom_smooth() +
  facet_grid(sex ~ race)
```

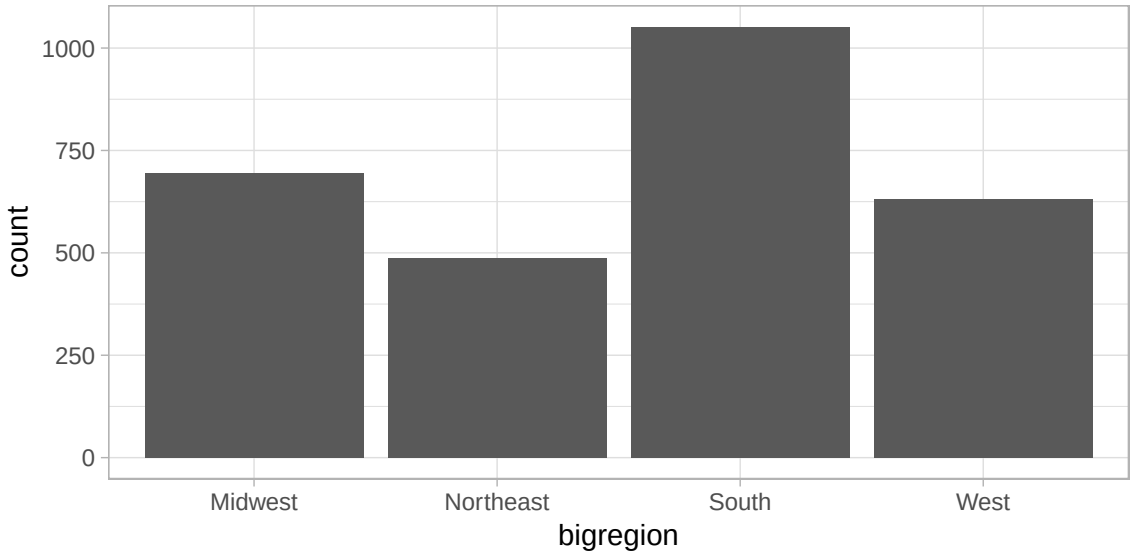
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'

## Warning: Removed 18 rows containing non-finite values (stat_smooth).

## Warning: Removed 18 rows containing missing values (geom_point).
```



Geom (Geometrias) pueden transformar los datos.

- cada `geom_` tiene su funcion de `stat_` asociado que utiliza por defecto que muchas veces no es obvio de base consideremos `geom_bar()`

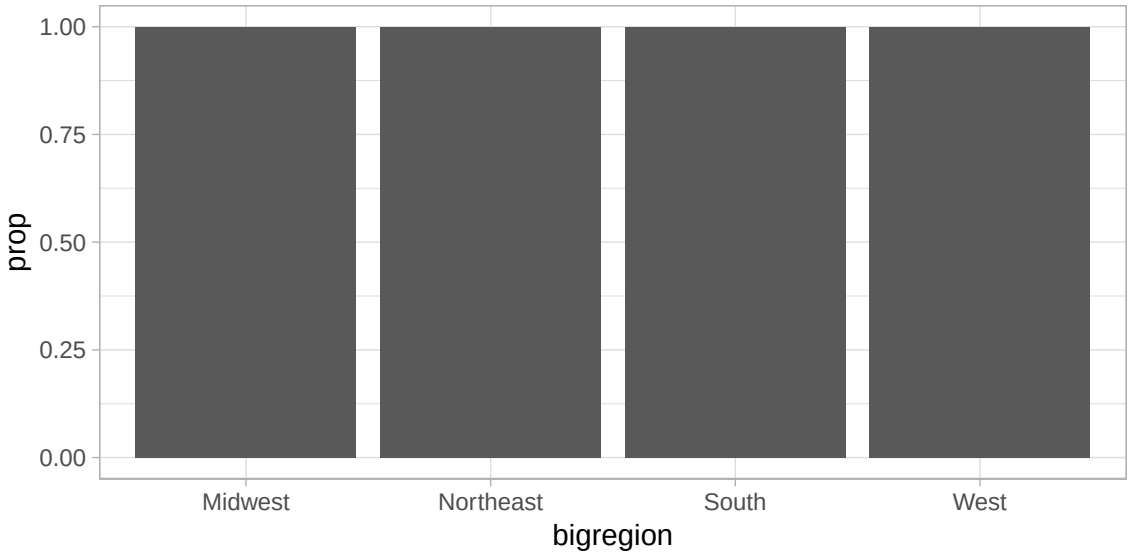
```
ggplot(data = base_gss_sm, aes(x = bigregion)) +
  geom_bar()
```



Como ven en el ejemplo solo provisionamos una estetica que fue el eje-x con `bigregion`, sin embargo el`geom_bar` hizo un conteo de la variable para graficar el diagrama de barra.

Si queremos ver las proporciones relativas podemos indicarle que lo mapee en el eje-y.

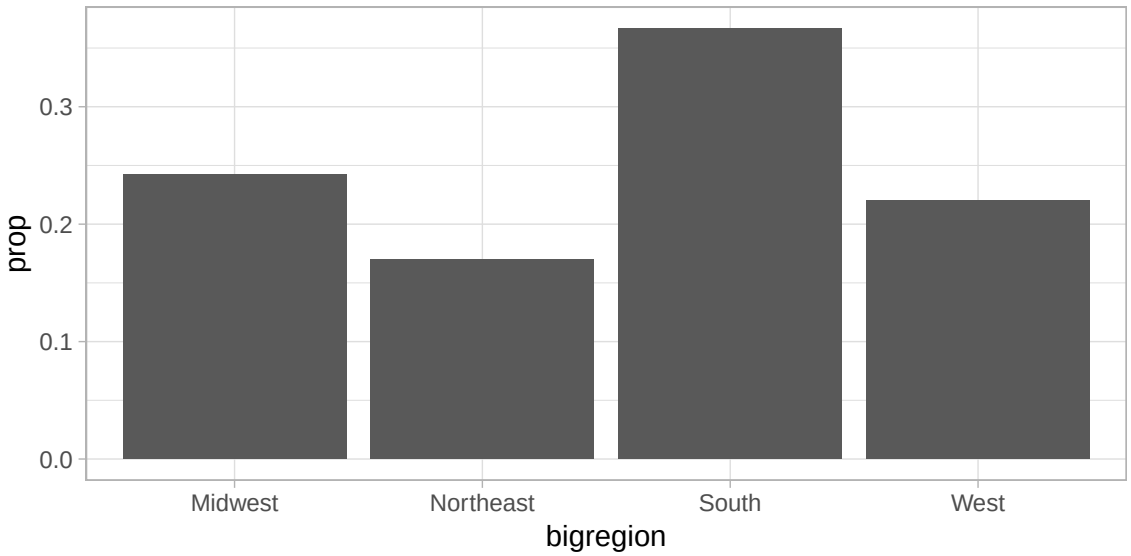
```
ggplot(data = base_gss_sm, aes(x = bigregion)) +
  geom_bar(aes(y = ..prop..))
```



Aca la estadistica relevante fue `..prop...`

Sin embargo el grafico sigue como no informativo, ya no tenemos un conteo en eje-y pero proporciones con valor de 1 (suma). En este caso necesitamos decirle a ggplot que ignore las categorias-x cuando calcule el denominador de la proporcion y que use el numero total de observaciones. Esto lo logramos con `group = 1`

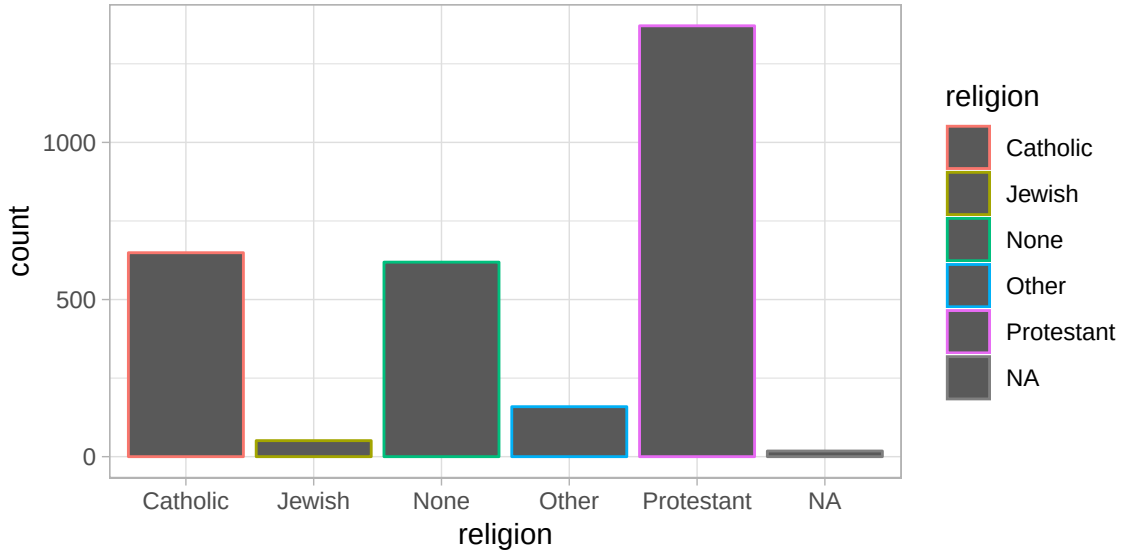
```
ggplot(data = base_gss_sm, aes(x = bigregion)) +
  geom_bar(aes(y = ..prop.., group = 1))
```



Vamos a explorar otra variable que es la religion: `religion`.

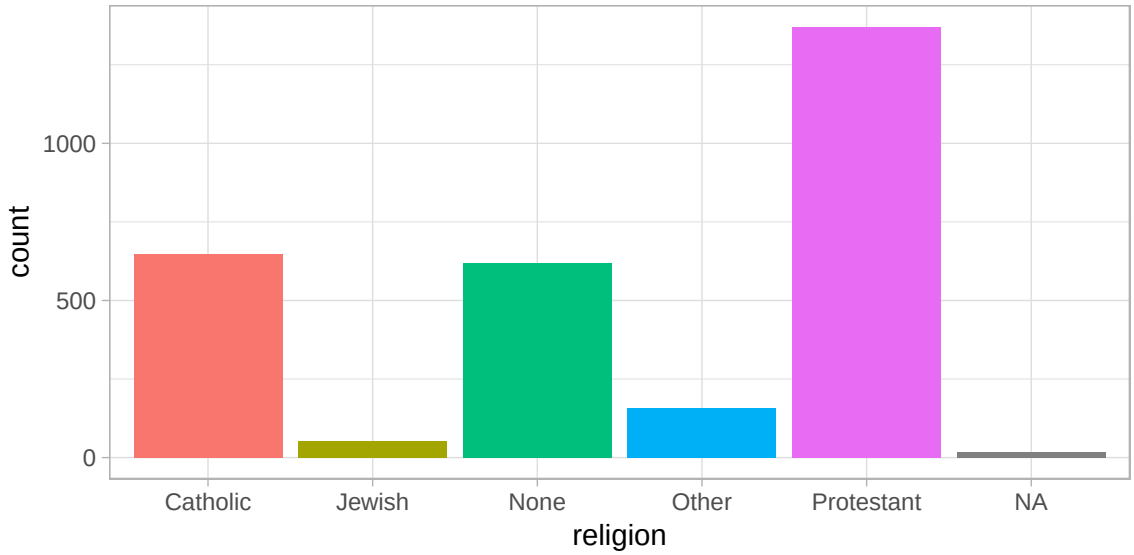
Generen un grafico de barras de la religion mapeando religion al eje-x y color.

```
ggplot(data = base_gss_sm,
           mapping = aes(x = religion, color = religion)) +
geom_bar()
```



Vemos que color solo le da color al borde para rellenar las barra tenemos que usar fill.

```
ggplot(data = base_gss_sm,
           mapping = aes(x = religion, fill = religion)) +
  geom_bar() +
  guides(fill = FALSE)
```

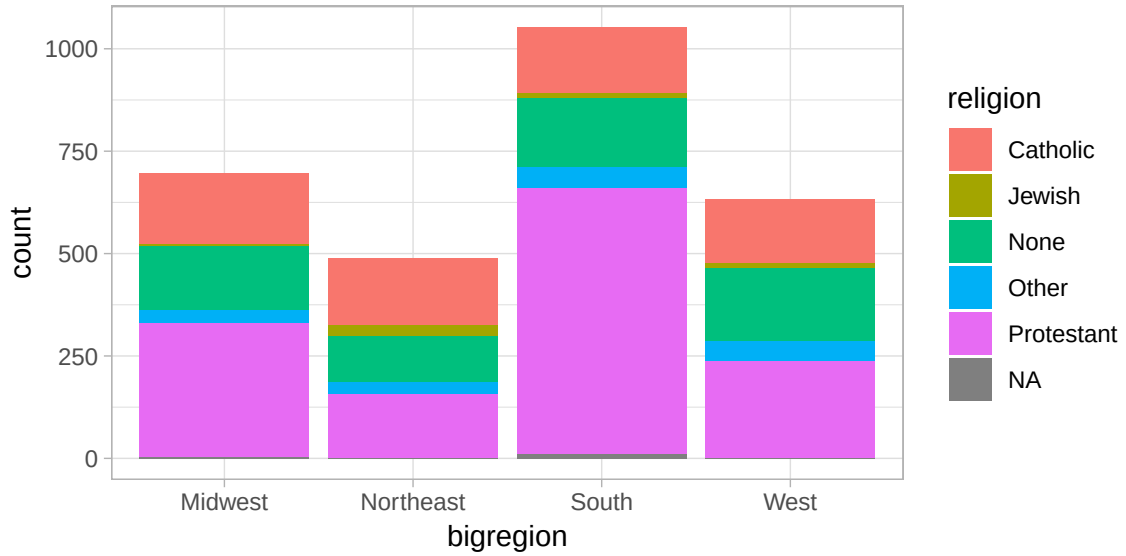


Con `guides` quitamos la leyende en los graficos de barras ya que este es redundante para las categoria. Recuerden la operacion de `show.legend = FALSE` donde iria esta? creen que funciona?

Graficos de frecuencias.

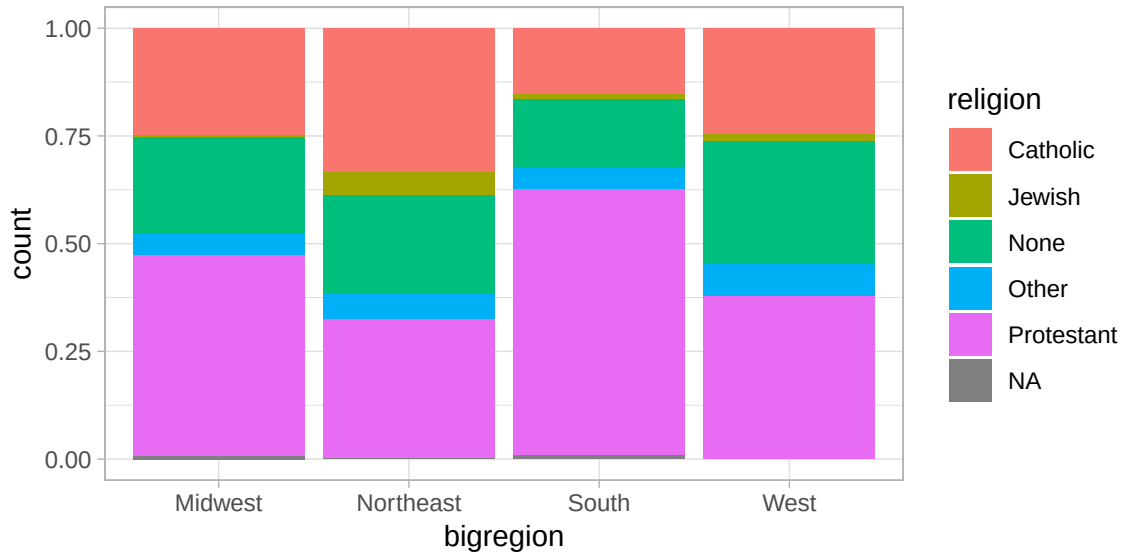
Vamos hacer un grafico de barras para religion por region

```
ggplot(data = base_gss_sm,
           mapping = aes(x = bigregion, fill = religion)) +
  geom_bar()
```



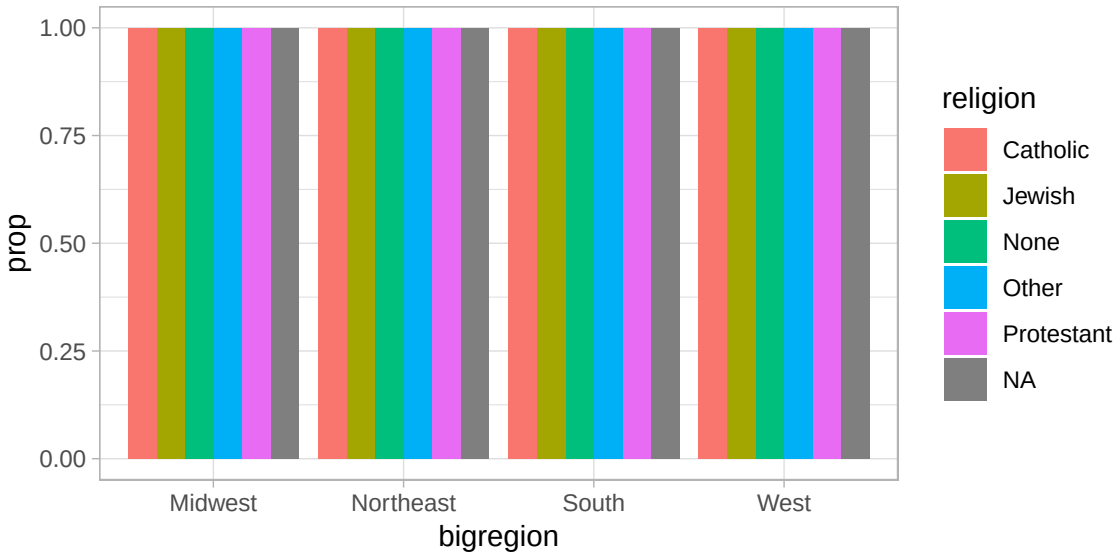
Este grafico no nos ayuda mucho queremos ver es mejor ver las proporciones. puedent utilizar el fill como posicion. (position = "fill")

```
ggplot(data = base_gss_sm,
       mapping = aes(x = bigregion, fill = religion)) +
  geom_bar(position = "fill")
```



Sin embargo esto o nos ayuda mucho ya que queremos ver los graficos al lado de otro

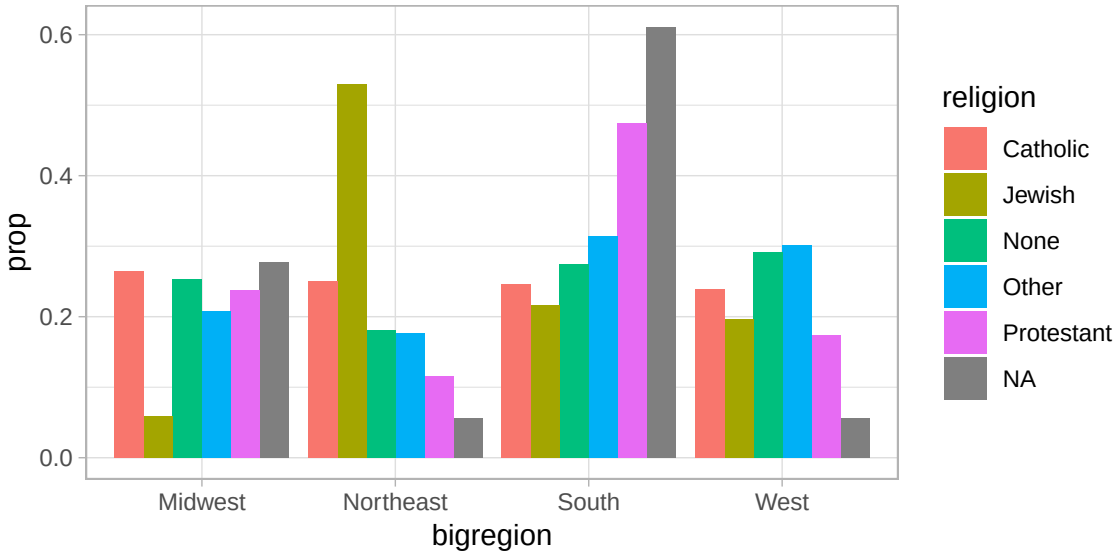
```
ggplot(data = base_gss_sm,
           mapping = aes(x = bigregion, fill = religion)) +
geom_bar(position = "dodge",
         mapping = aes(y = ..prop..))
```



Tampoco funciona

Arreglen la grafica para lograr lo que queremos.

```
ggplot(data = base_gss_sm,
       mapping = aes(x = bigregion, fill = religion)) +
  geom_bar(position = "dodge",
           mapping = aes(y = ..prop.., group = religion))
```



Sin embargo podemos dejar de forzar ggplot que nos de ese grafico y utilizar faceta

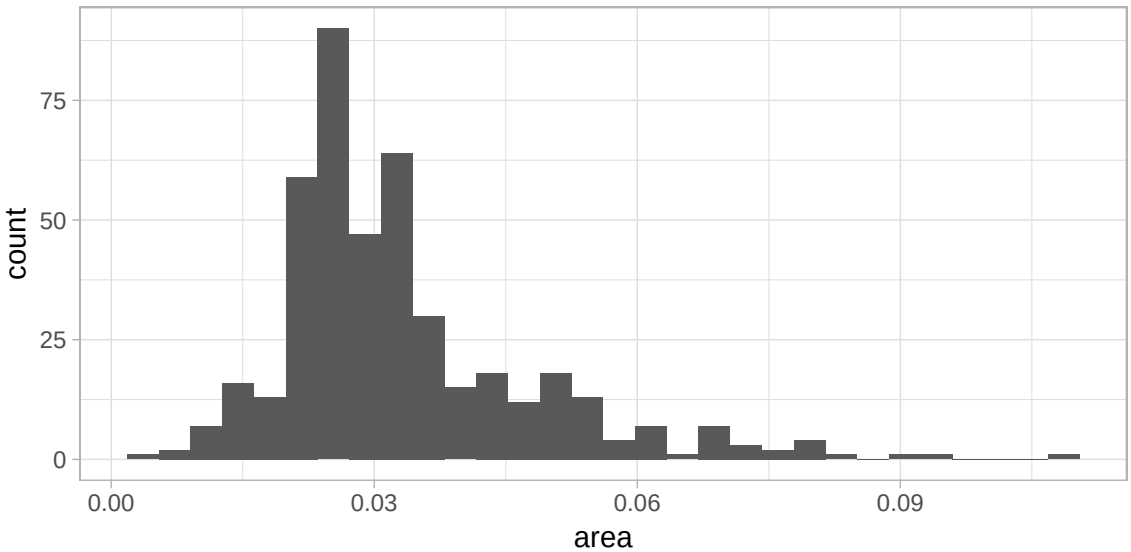
```
ggplot(data = base_gss_sm,
            mapping = aes(x = religion, fill = religion)) +
  geom_bar() +
  facet_wrap(~ bigregion)
```

Histogramas y graficos de densidad.

El histograma es una forma de transformar variables continuas "rompiendolas" en grupos pequeños, llamados "bins". Para la siguiente utilizaremos los datos de `midwest` que esta en el paquete de `ggplot2`.

```
ggplot(data = midwest,
       mapping = aes(x = area)) +
  geom_histogram()
```

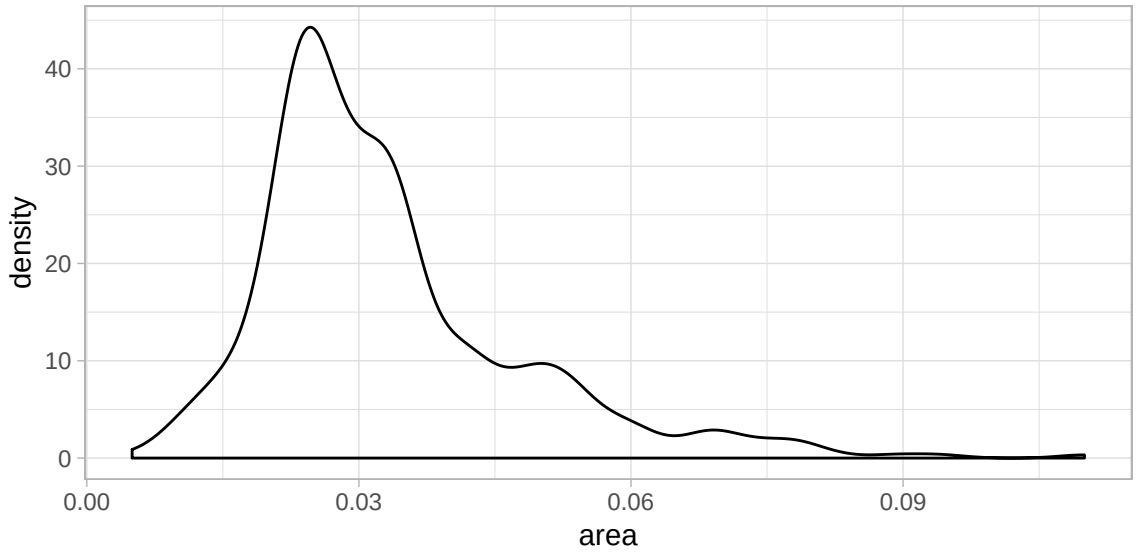
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Como vimos anteriormente aca una nueva variable `count` aparece en el eje-y

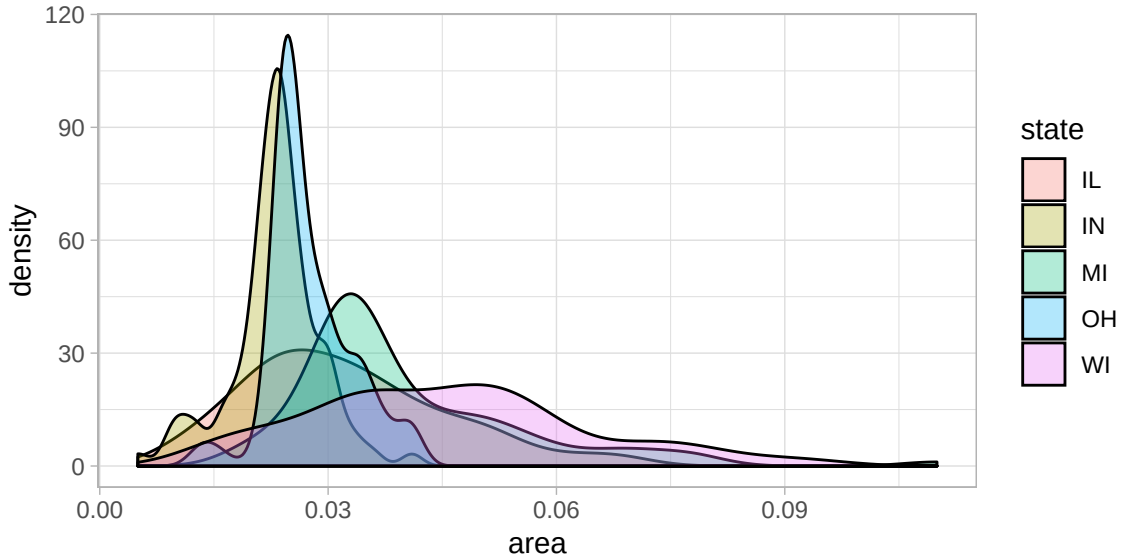
Para un grafico de densidad usamos `geom_density`

```
ggplot(data = midwest,
       mapping = aes(x = area)) +
  geom_density()
```



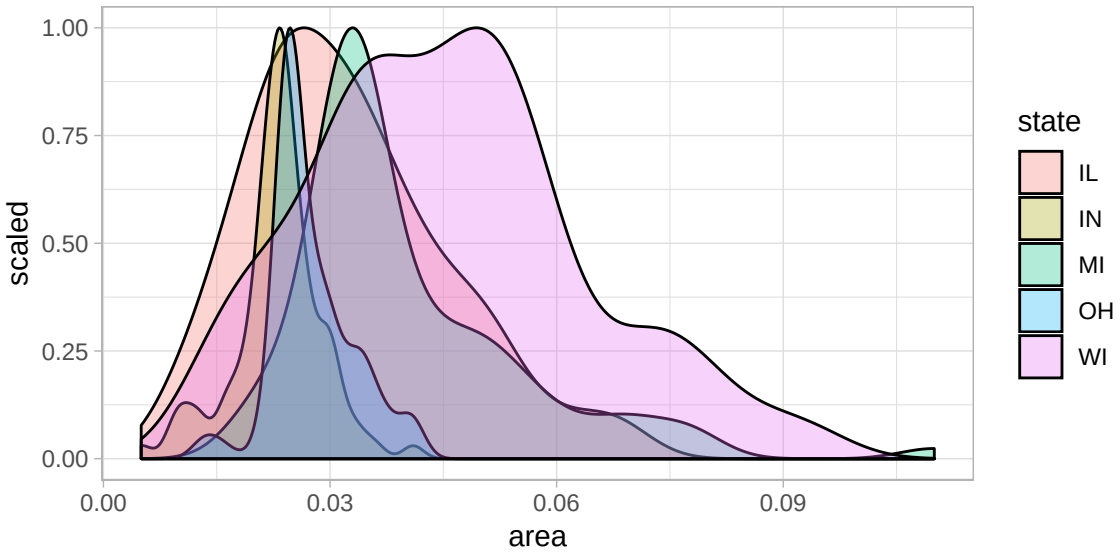
Podemos incluir color y relleno

```
ggplot(data = midwest,
       mapping = aes(x = area, fill = state)) +
  geom_density(alpha = 0.3)
```



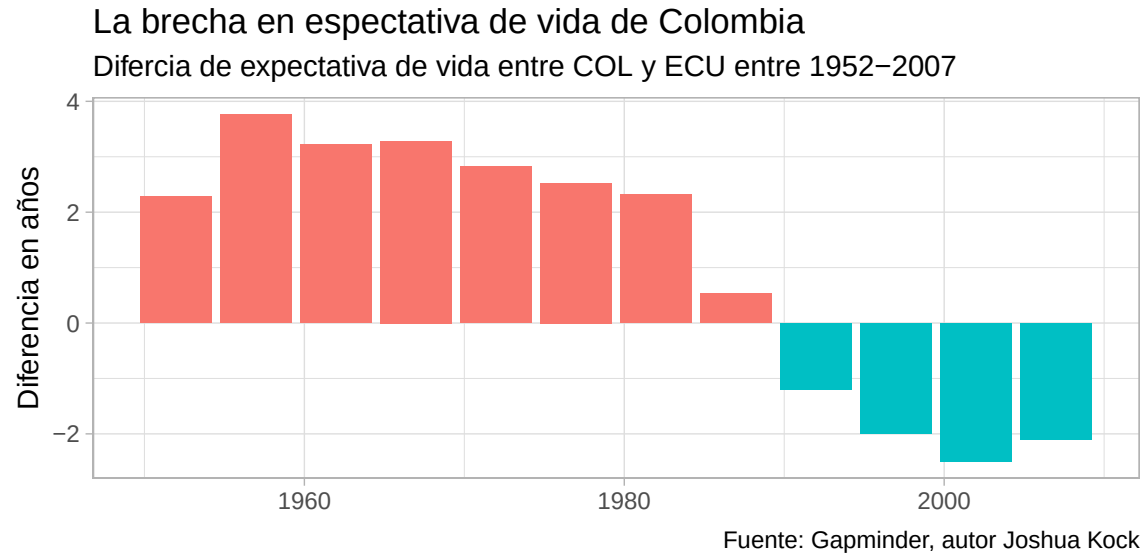
Igual como vimos en los graficos de barra podemos ajustar la estadistica `stat_` para obtener una "proporcion" o en este caso una escala.

```
ggplot(data = midwest,
       mapping = aes(x = area, fill = state)) +
  geom_density(alpha = 0.3, aes(y = ..scaled..))
```



Generar la siguiente grafica con los datos de gapminder

```
gapminder %>%
  filter(country == "Colombia" | country == "Ecuador") %>%
  select(country, year, lifeExp) %>%
  spread(country, lifeExp) %>%
  mutate(dif_life_exp = Colombia - Ecuador,
         hi_low = if_else(dif_life_exp < 0, "bajo", "alto")
  ) %>%
  ggplot(aes(x = year, y = dif_life_exp, fill = hi_low)) +
  geom_col(show.legend = FALSE) +
  labs(
    title = "La brecha en espectativa de vida de Colombia",
    x = NULL,
    y = "Diferencia en años",
    subtitle = "Difercia de expectativa de vida entre COL y ECU entre 1952-2007",
    caption = "Fuente: Gapminder, autor Joshua Kock"
  )
```



Recursos:

Para mas informacion puedes consultar estos recursos:

http://www.sthda.com/english/wiki/be-awesome-in-ggplot2-a-practical-guide-to-be-highly-effective-r-software-and-data-vis

https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf

https://www.tandfonline.com/doi/abs/10.1198/jcgs.2009.07098

http://www.ggplot2-exts.org/gallery/