Sección 3

Entendiendo y presentando datos con Gapminder Foundation

- Entender como Hans Rosling y Gapminder Foundation utilizan técnicas efectivas de visualización de datos para comunicar y transmitir tendencias.
- Poder aplicar las herramientas vistas en ggplot2 para responder preguntas con los datos
- Entender como escalas fijas en gráficos simplifican comparaciones
- Poder modificar gráficos para mejorar la visualización de los datos

En esta sección utilizaremos datos para responder a estas dos preguntas

1) ¿Es justa la caracterización del mundo diciendo que se encuentra dividido entre naciones ricas occidentales y otras en vías de desarrollo siendo estas África, Asia, y América latina?
2) ¿Se ha agravado en los últimos 40 años la desigualdad de ingresos a través de los diferentes países?

Análisis de datos utilizando dataset Gapminder

- Gapminder charlas Ted:
  - Las mejores estadísticas que has visto
  - Nuevos descubrimientos en la pobreza

La visualización de datos puede utilizarse para romper mitos y educar contradiciendo historias sensacionalistas o fuera de fechas

El proyecto que utilizaremos en la librería dslabs:

```
library(dslabs)
data(gapminder)
```

Test de Hans Rosling

1- Sri Lanka o Turquía

2- Polonia o Corea del sur

3- Malasia o Rusia

4- Pakistán o Vietnam

5- Tailandia o Sudáfrica

- Para cada par de países en el 2015
  - ¿Que país piensas que tiene la tasa de mortalidad infantil más alta?
  - Cuales países piensas que son similares?

Test de Hans Rosling

1- Sri Lanka o Turquía

2- Polonia o Corea del sur

3- Malasia o Rusia

4- Pakistán o Vietnam

5- Tailandia o Sudáfrica

- Sin datos responderíamos que los países no europeos tienen una tasa de mortalidad infantil más alta
- Los países considerados como en vías de desarrollos pensaríamos que tienen una tasa alta de moralidad similar.

Utilizando los datos de Gapminder

```
gapminder %>% filter(year == 2015 & country %in% c("Sri Lanka", "Turkey")) %>% select(country, infant_mortality)
```

```
    country infant_mortality
1 Sri Lanka              8.4
2    Turkey             11.6
```

country	infant mortality	country	infant mortality
Sri Lanka	8.4	Turkey	11.6
Poland	4.5	South Korea	2.9
Malaysia	6.0	Russia	8.2
Pakistan	65.8	Vietnam	17.3
Thailand	10.5	South Africa	33.6

Expectativa de vida y fertilidad - Concepciones

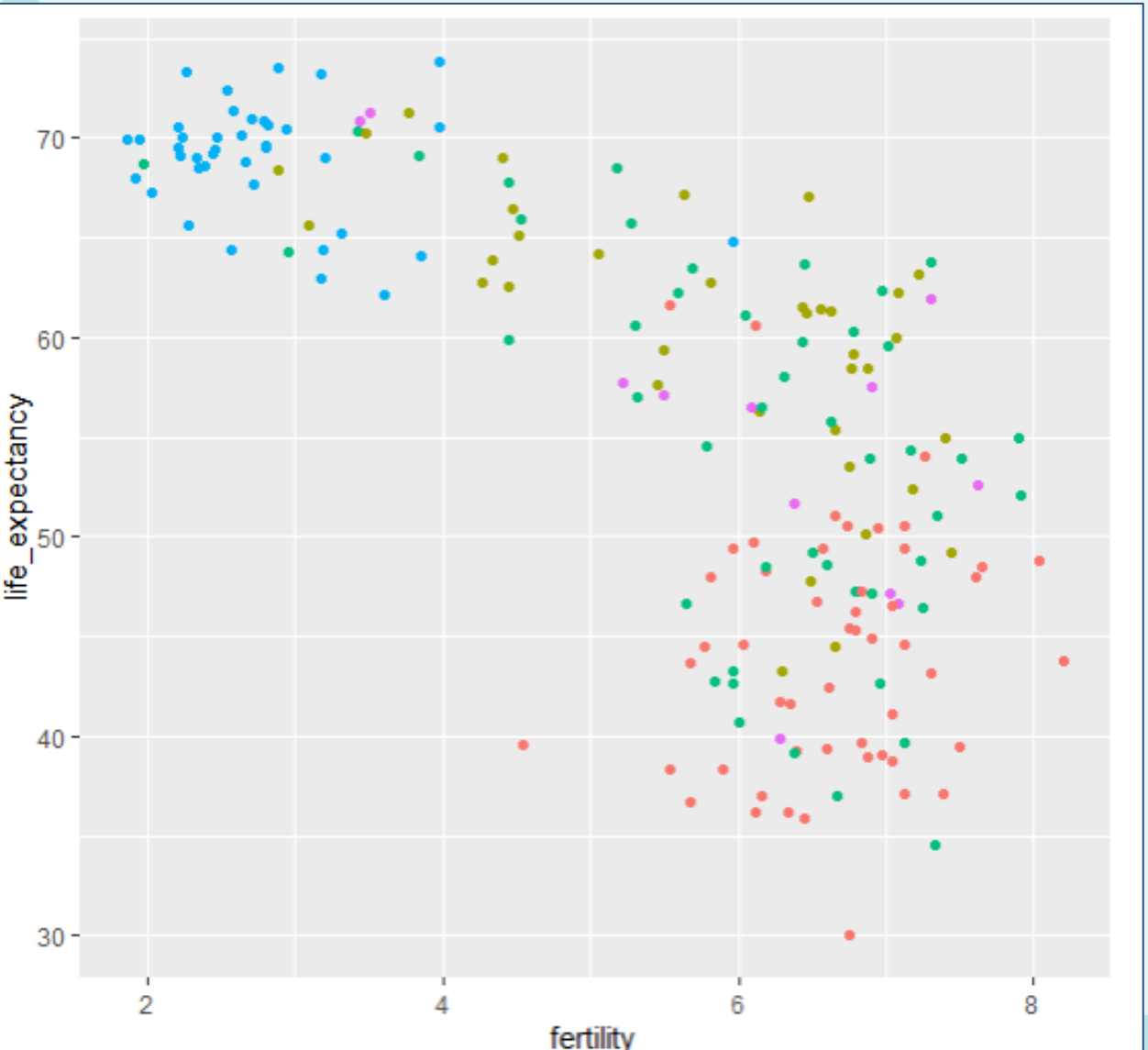
1) Mundo occidental - Europa del Oeste y américa del norte

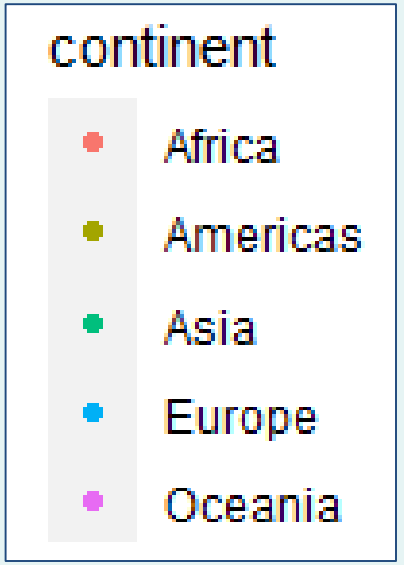
   Larga esperanza de vida y familias pequeñas

1) Mundo en vías de desarrollo (África, Asia y América Latina)

   Esperanza de vida menor y familias grandes

```
ds_theme_set()
filter(gapminder, year == 1962) %>%
ggplot(aes(fertility, life_expectancy, color = continent)) +  geom_point()
```

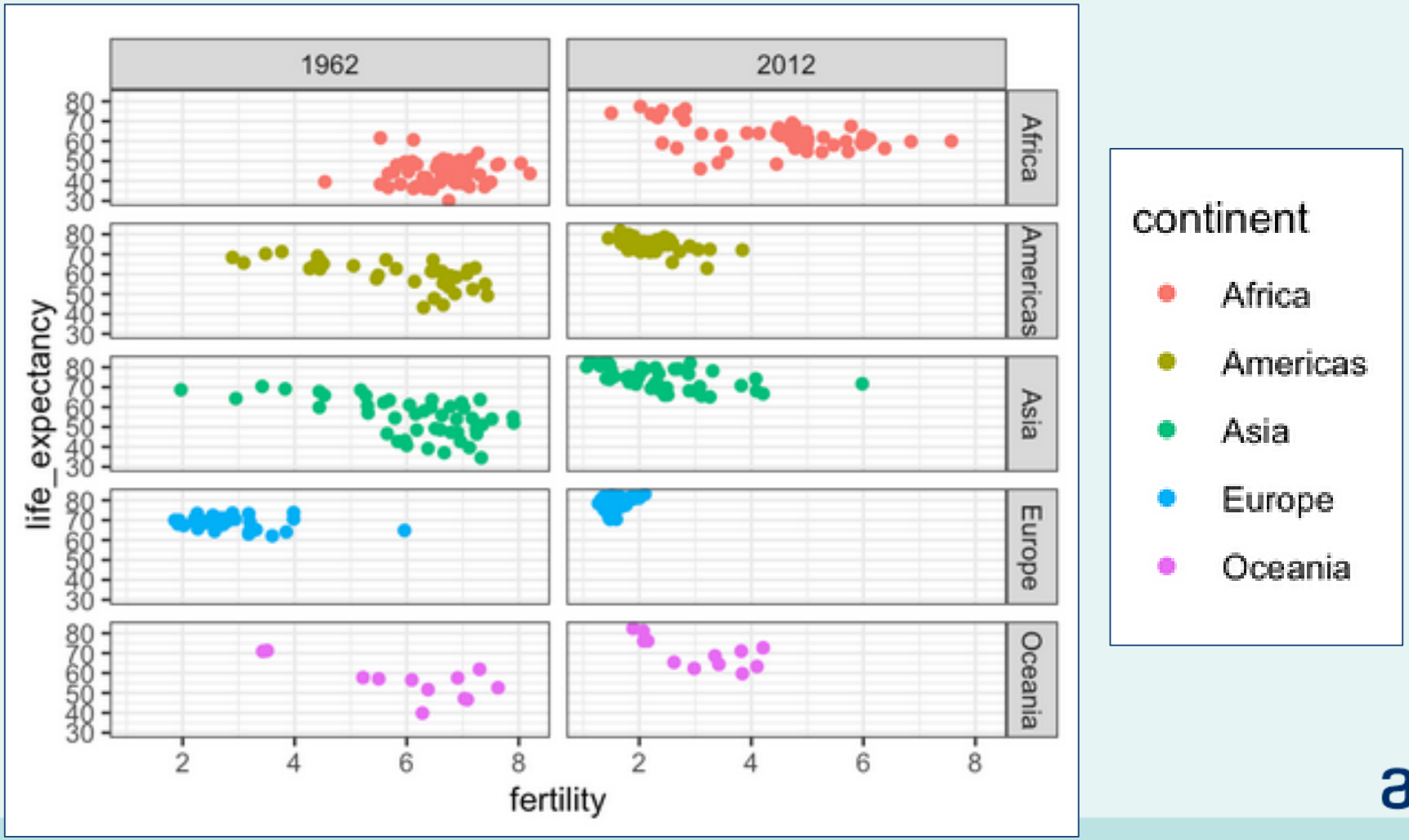




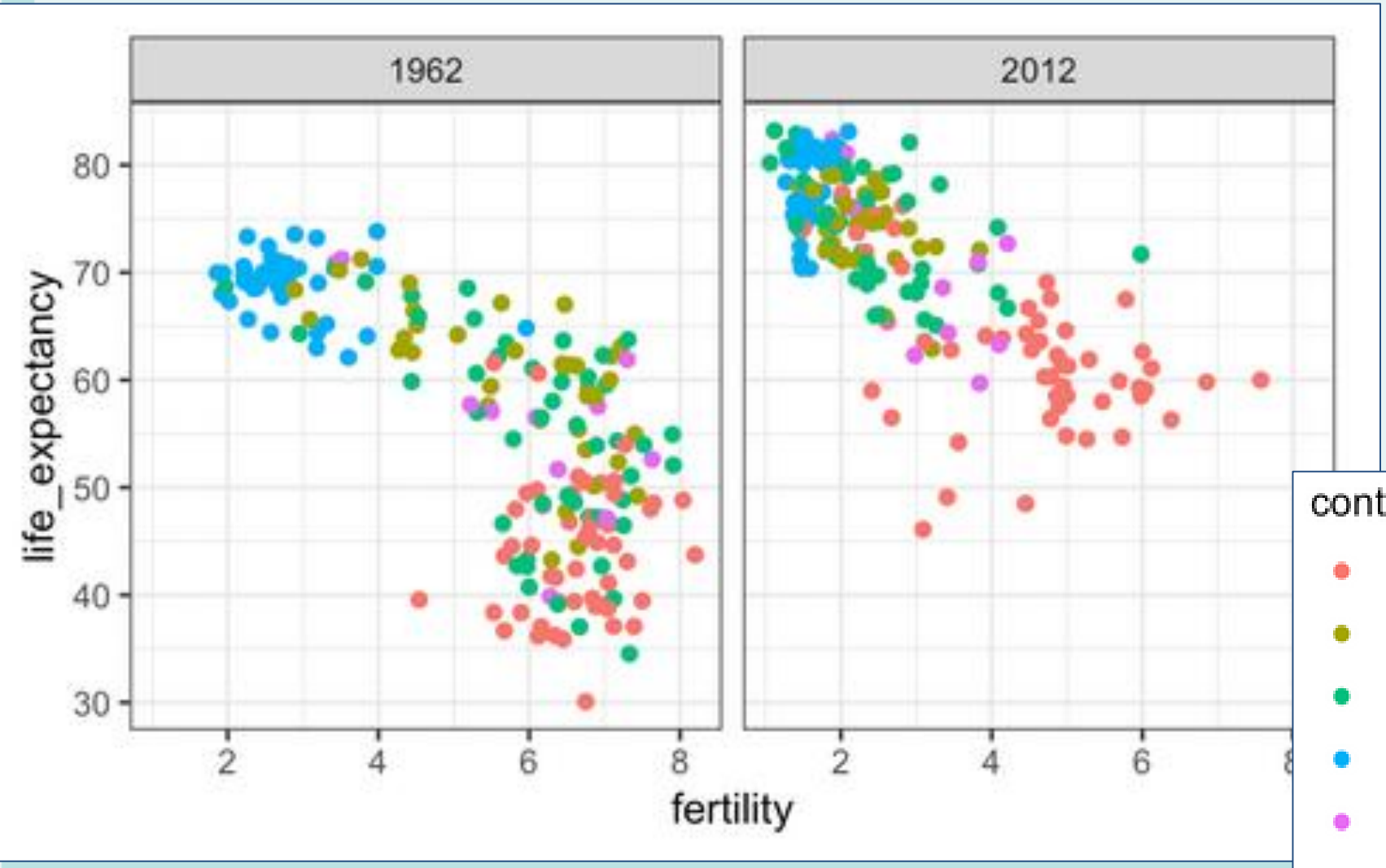
Pulir, filtrar datos - Faceting

- Cuando queremos pulir o comparar gráficos de acuerdo a variables
- **facet_grid()** nos permite manejar varias variables, una en filas y otra en columnas.
- **facet_wrap()** permite manipular una variable para visualizar una serie de gráficos en una tabla con dimensiones legibles, adaptándola al display.
- Cuando pulimos datos las escalas quedan fijas para todos los gráficos.

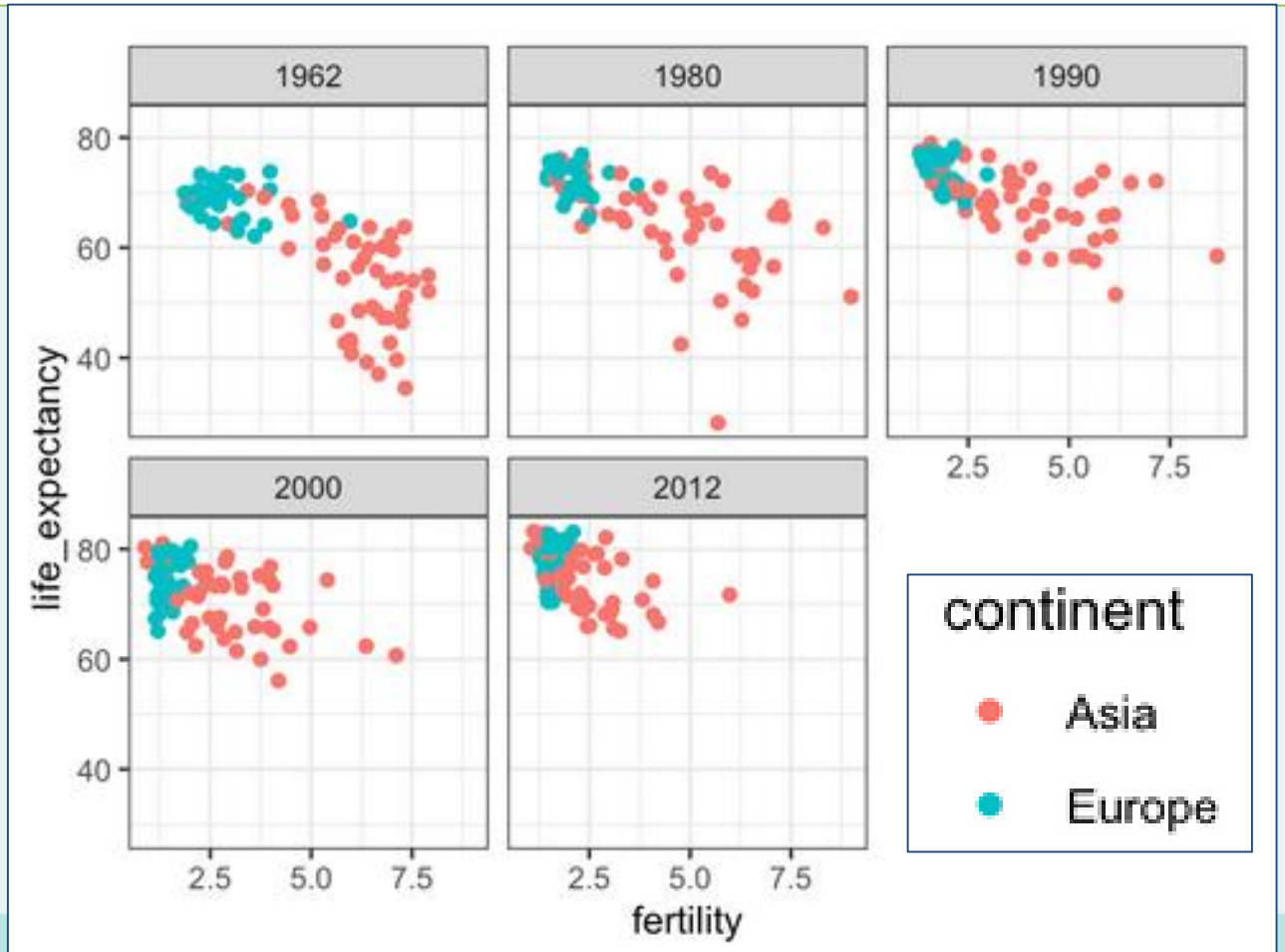
```
filter(gapminder, year%in%c(1962, 2012)) %>%
  ggplot(aes(fertility, life_expectancy, col = continent)) + geom_point() +
  facet_grid(continent~year)
```



```
filter(gapminder, year%in%c(1962, 2012)) %>%
  ggplot(aes(fertility, life_expectancy, col = continent)) +
  geom_point() +
  facet_grid(. ~ year)
```



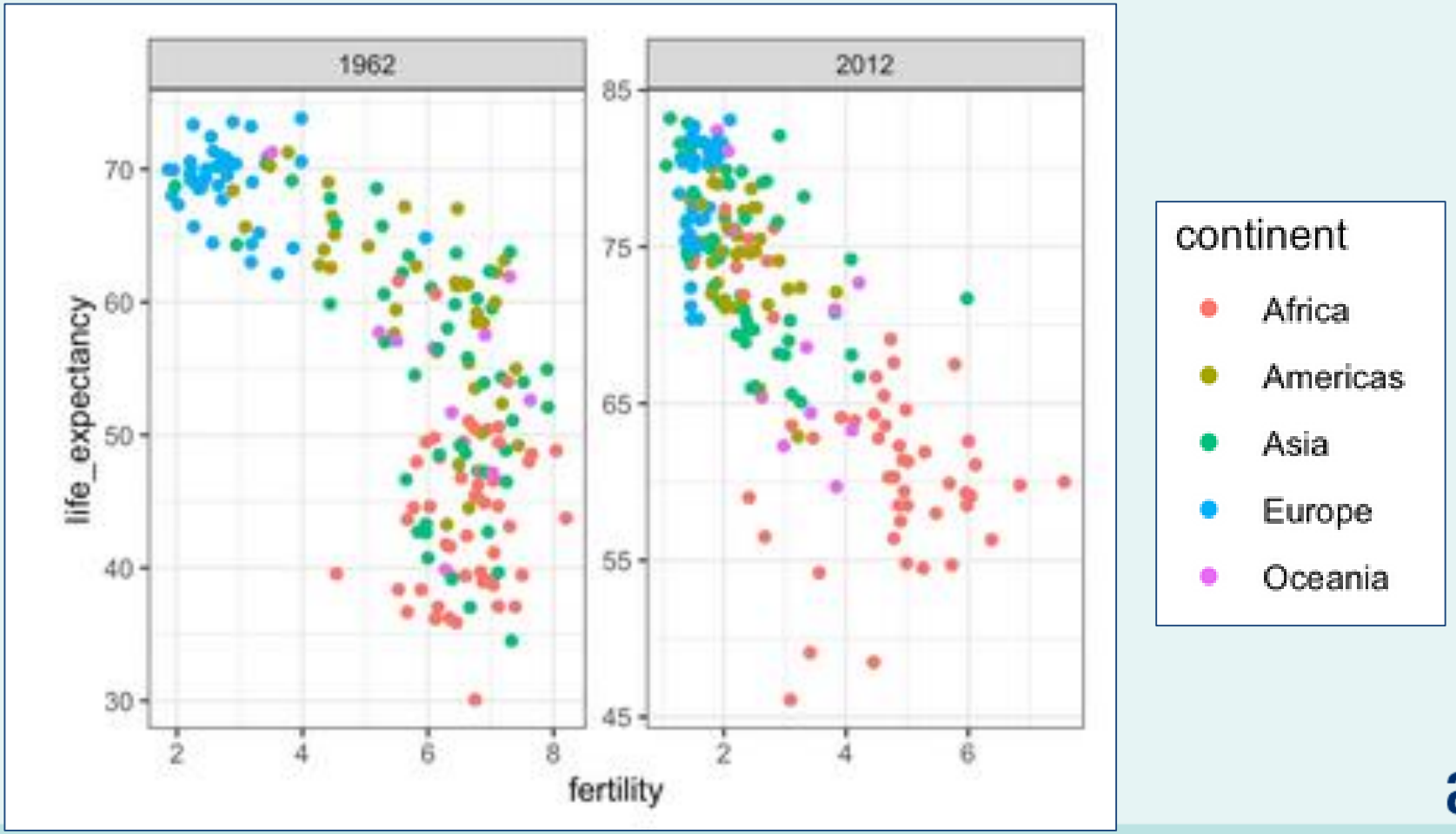
```
years <- c(1962, 1980, 1990, 2000, 2012)
continents <- c("Europe", "Asia")
gapminder %>% filter(year %in% years & continent %in% continents) %>%
  ggplot( aes(fertility, life_expectancy, col = continent)) + geom_point() +
  facet_wrap( vars(year) )
```



Ejes y escalas

- Cuando graficamos sin utilizar la función facet las escalas y rangos se determinan automáticamente por los datos del gráfico.
- Utilizando facet el rango y escalas de datos se determina por todos los gráficos a realizar y se fija en uno único para todos los elementos.

```
filter(gapminder, year%in%c(1962, 2012)) %>%
  ggplot(aes(fertility, life_expectancy, col = continent)) +
geom_point() + facet_wrap(. ~ year, scales = "free")
```

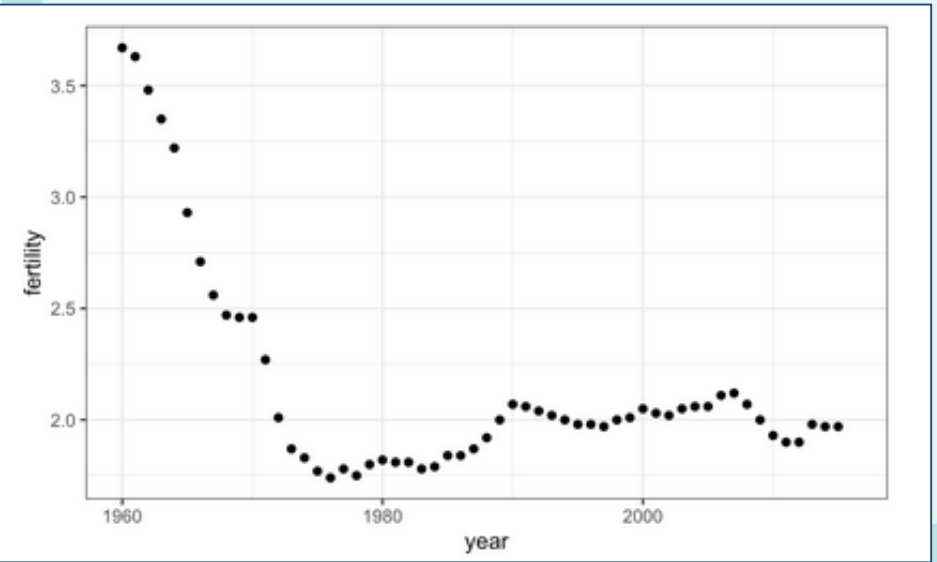


Gráficos de tiempo

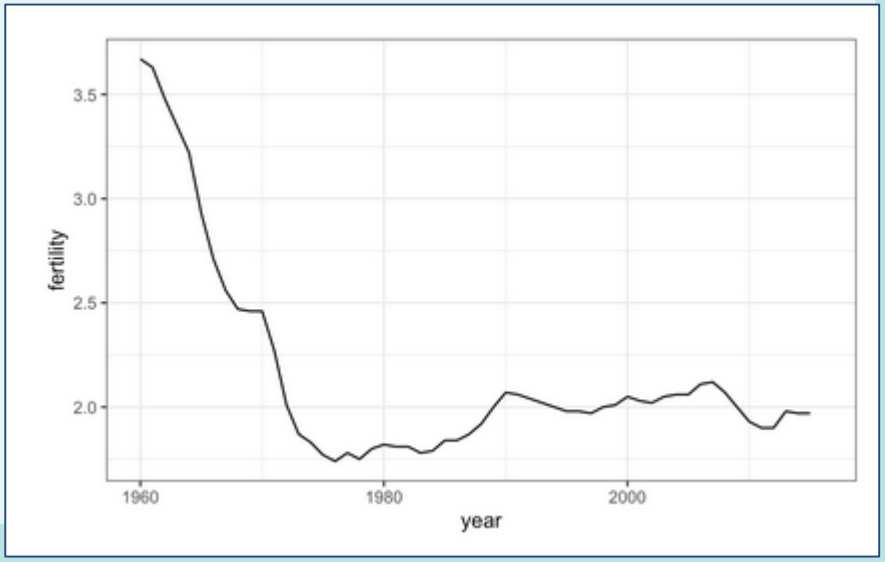
- Tienen el tiempo en el eje x y una salida o medida en el eje y

```
gapminder %>%
  filter(country == "United States") %>% ggplot(aes(year, fertility)) +
```

geom_point()



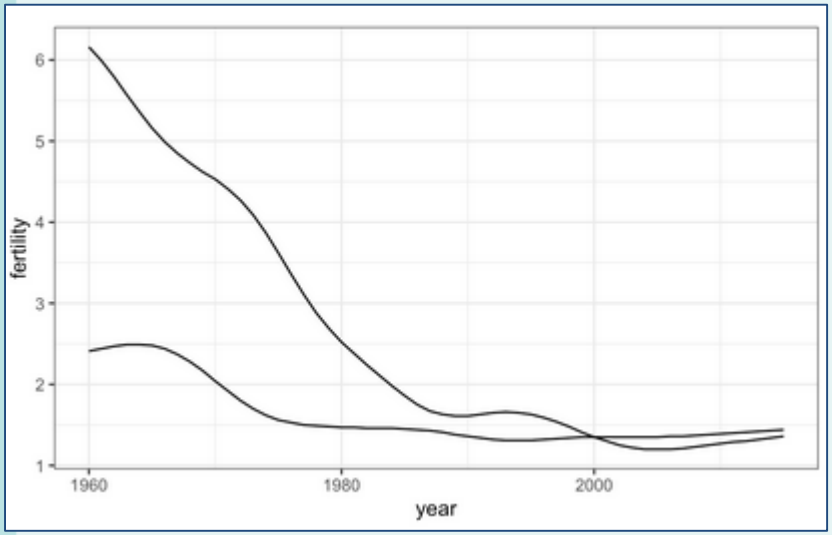
geom_line()

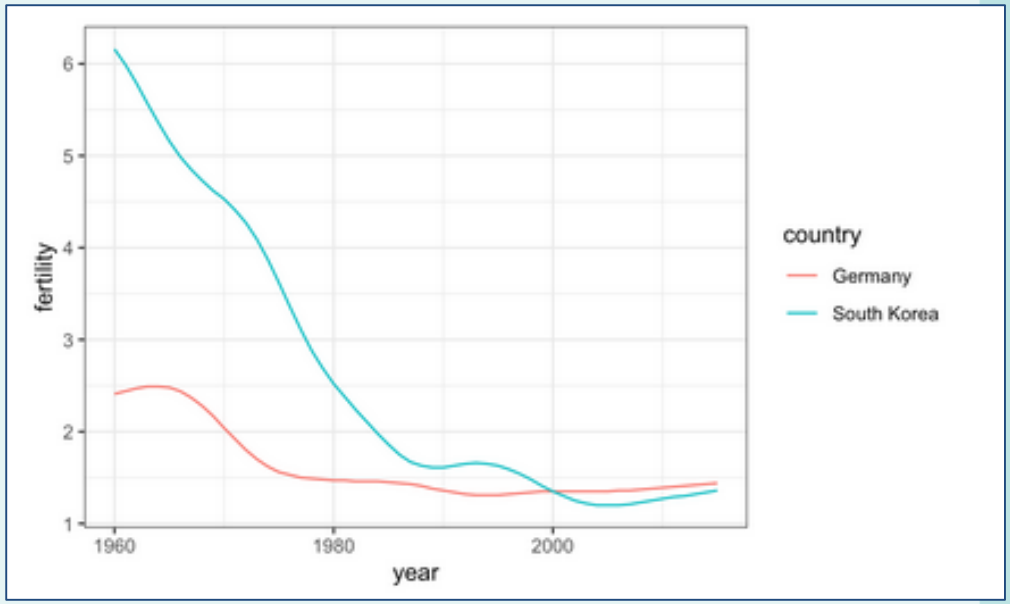


Comparamos índices en dos países

```
countries <- c("South Korea","Germany")
gapminder %>% filter(country %in% countries) %>%
    ggplot(aes(year, fertility, group = country)) +
    geom_line()
```

col = country

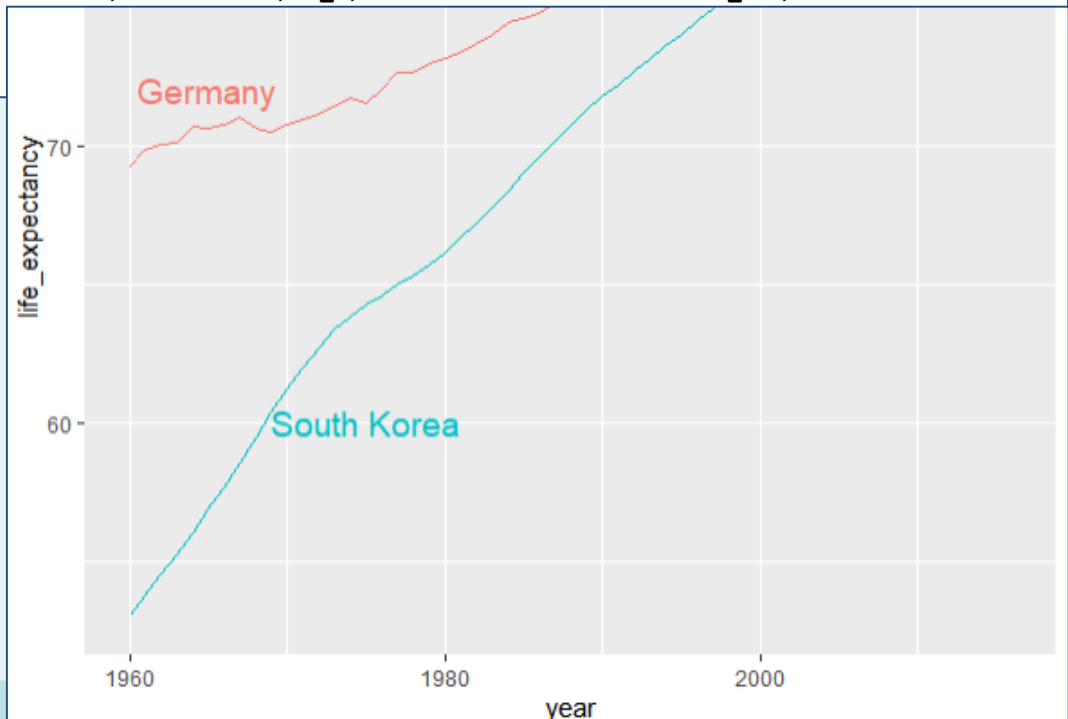




Textos en gráficos - Labels

- Definimos un data frame con textos y posiciones

```
> countries <- c("South Korea","Germany")
> labels <- data.frame(country = countries, x = c(1975,1965), y = c(60,72))
> gapminder %>% filter(country %in% countries) %>%
ggplot(aes(year, life_expectancy, col = country)) +
+ geom_line() + geom_text(data = labels, aes(x, y, label = country), size = 5)
+ theme(legend.position = "none")
```



Tipos de datos

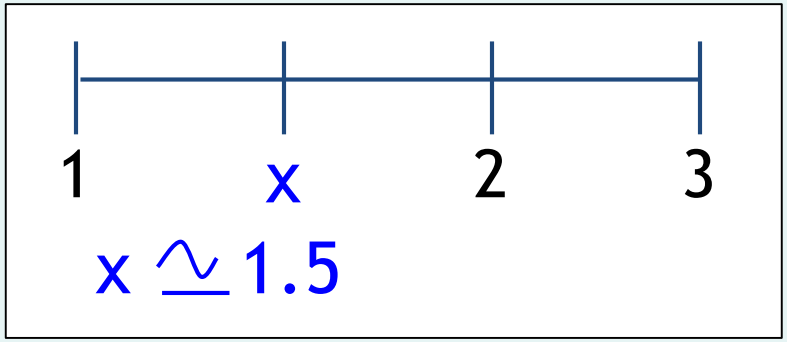
- En exploración y análisis de datos los ejes logarítmicos son muy utilizados:
  - Log (log base 2 y log base 10) son más visualizables:

    2^2,2^3,2^4,… o 10^2,10^3,…
  - Logaritmo natural ln es menos intuitivo y más difícil de interpretar

    e^2, e^3,….

Transformar datos o escalas

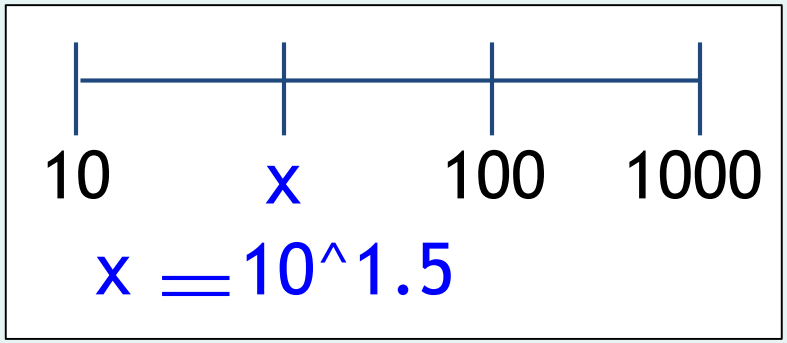
- Podemos realizar log de los valores antes de graficarlos e interpretar fácilmente:



Vemos los valores originales en los ejes

- Podemos utilizar ejes logarítmicos.

```
scale_x_continuous(trans = "log2")
scale_x_log10()
```



Los valores originales se ven en el gráfico

Re ordenar datos -factores

- Por defecto R ordena los factores alfabéticamente

nombres	notas	edades
Ana	D	19
Valeria	MB	20
Cecilia	B	11
Maria	B	9

```
> class(alumnos$notas)
[1] "factor"
> alumnos$notas
[1] D  MB B
Levels: B D MB
```

- Podemos definir un orden específico

```
>alumnos$notas<-ordered(alumnos$notas,levels=c("D","B","MB"))
> alumnos$notas
[1] D  MB B  B
Levels: D < B < MB
```

- Podemos ordenar de acuerdo a otro valor

```
> alumnos$notas <- reorder(alumnos$notas, alumnos$edades, FUN = mean)
> levels(alumnos$notas)
[1] "B"  "D"  "MB"
```

Datos categóricos

¿Existe económicamente una distinción entre el mundo occidental y el resto?

- Ingreso por día en dólares (pbi - gdp)

```
gapminder <- gapminder %>%
    mutate(dollars_per_day = gdp/population/365)
```

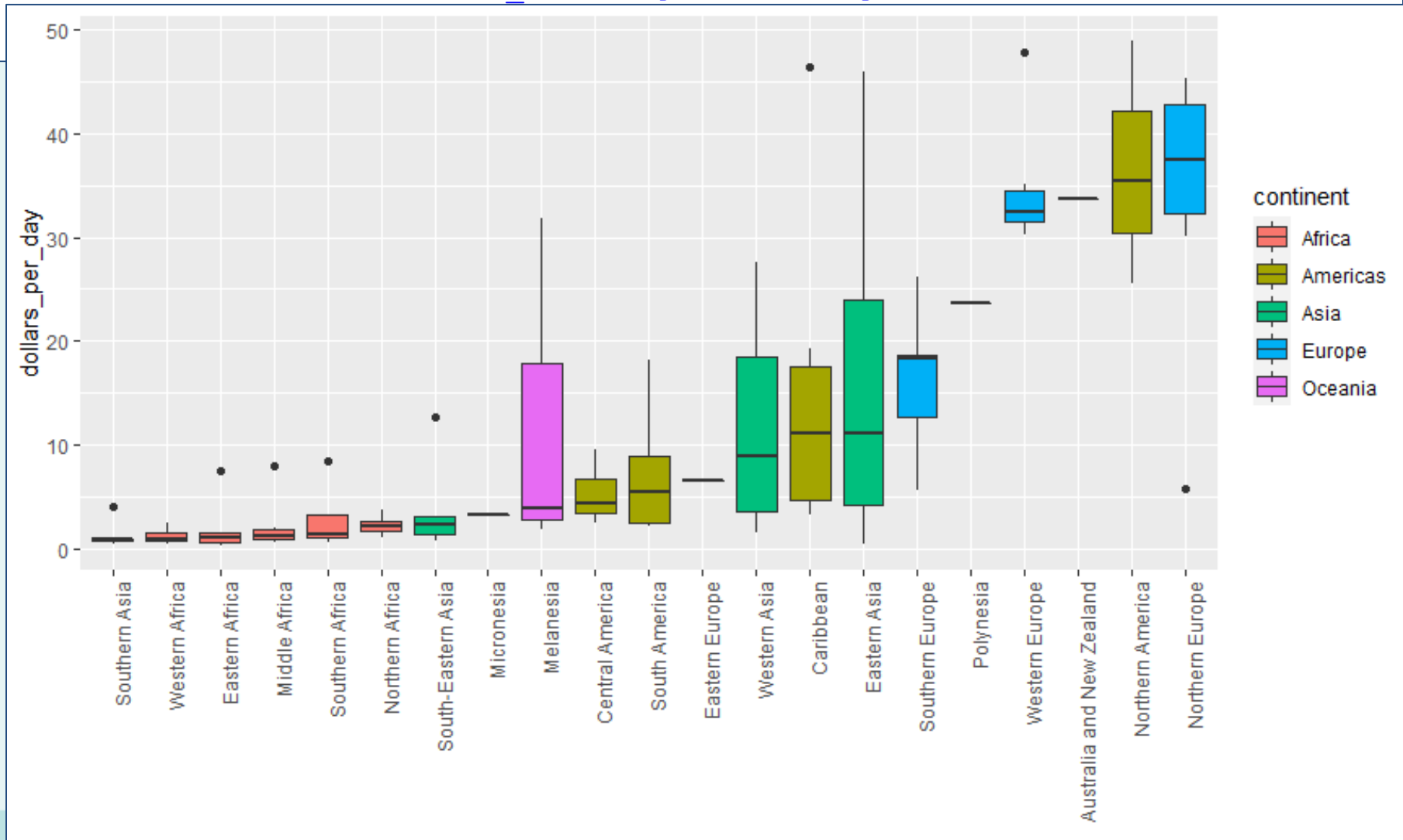
- Dividimos el mundo en regiones y ordenamos los datos

```
mutate(region = reorder(region, dollars_per_day, FUN = median))
```

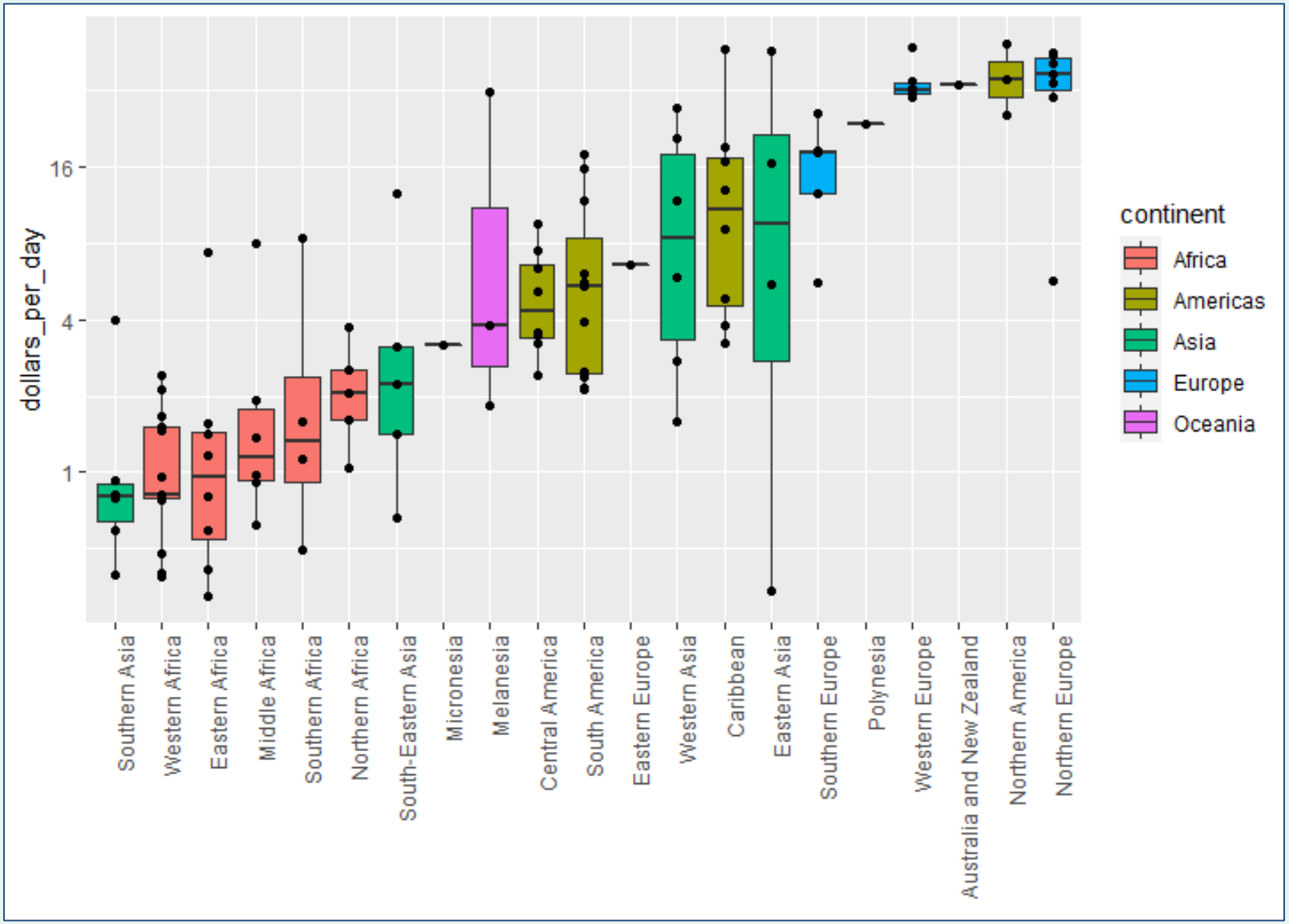
- Coloreamos gráfico por continente y movemos textos para evitar superposición

```
ggplot(aes(region, dollars_per_day, fill = continent)) +
geom_boxplot() +
theme(axis.text.x = element_text(angle = 90, hjust = 1))+ xlab("")
```

```
gapminder <- gapminder %>% mutate(dollars_per_day = gdp/population/365)
p <- gapminder %>% filter(year == 1970 & !is.na(gdp)) %>%
    mutate(region = reorder(region, dollars_per_day, FUN = median)) %>%
    ggplot(aes(region, dollars_per_day, fill = continent)) +
    geom_boxplot() +
    theme(axis.text.x = element_text(angle = 90, hjust = 1)) + xlab("")
p
```



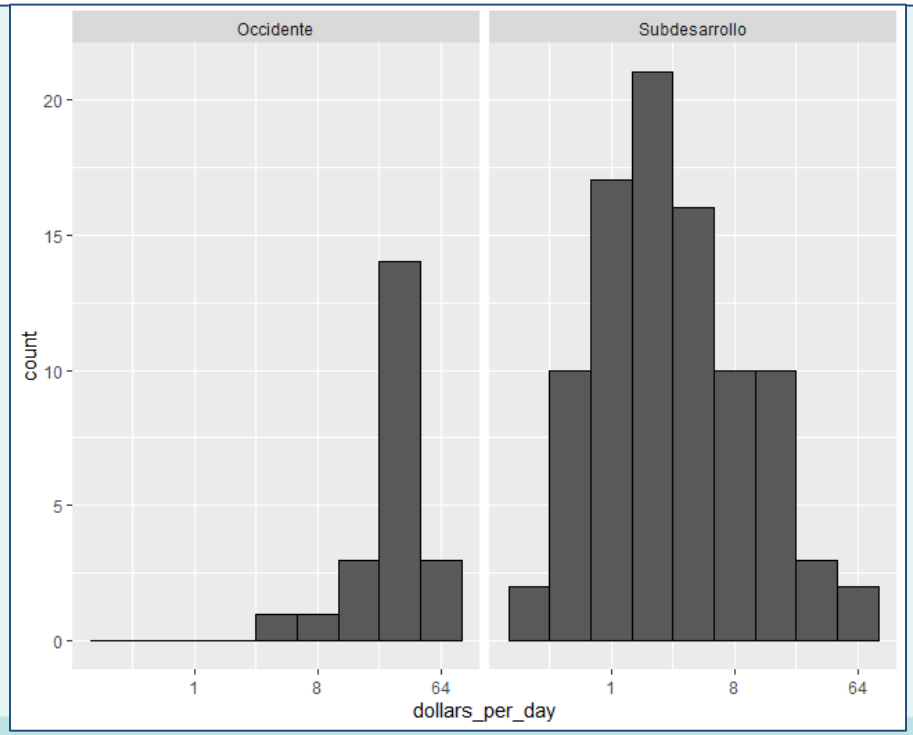
```
> p + scale_y_continuous(trans = "log2") + geom_point(show.legend= FALSE)
```



Comparación de distribuciones

- Clasificamos datos por grupo: Occidente-Subdesarrollados

```
past_year <- 1970
west <- c("Western Europe", "Northern Europe", "Southern Europe", "Northern
America", "Australia and New Zealand")
gapminder %>% filter(year == past_year & !is.na(gdp)) %>%
  mutate(group = ifelse(region %in% west, "Occidente", "Subdesarrollo")) %>%
  ggplot(aes(dollars_per_day)) + geom_histogram(binwidth = 1, color = "black")
+ scale_x_continuous(trans = "log2") + facet_grid(. ~ group)
```



Comparación de distribuciones

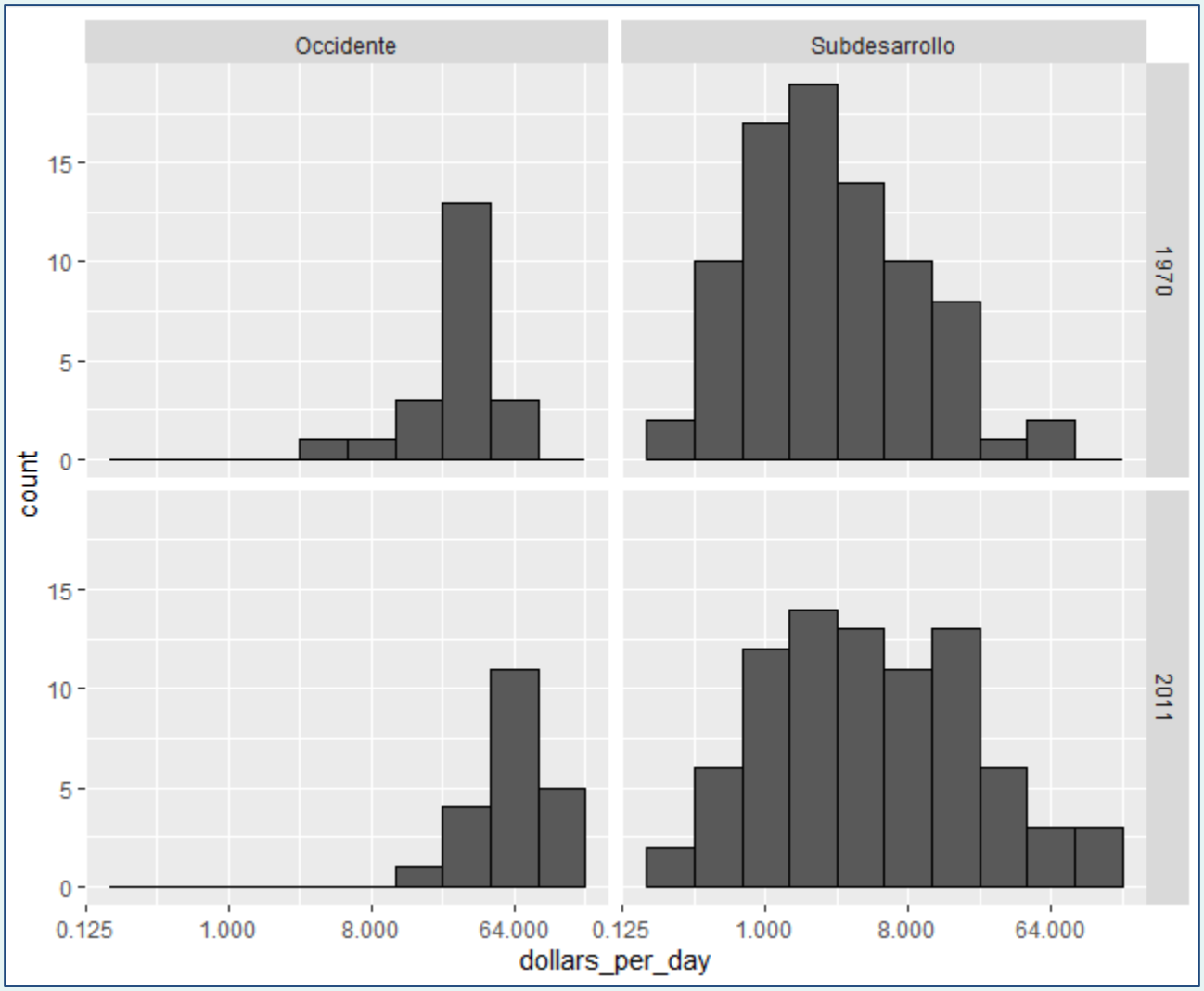
- Clasificamos datos por grupo y años
  - Definimos lista de países que se encuentren en ambos años

```
past_year <- 1970
present_year <- 2011
country_list_1 <- gapminder %>% filter(year == past_year &
!is.na(dollars_per_day)) %>% .$country
country_list_2 <- gapminder %>% filter(year == present_year &
!is.na(dollars_per_day)) %>% .$country
country_list <- intersect(country_list_1, country_list_2)
```

  - Graficamos incluyendo solo países con datos disponibles en esos años

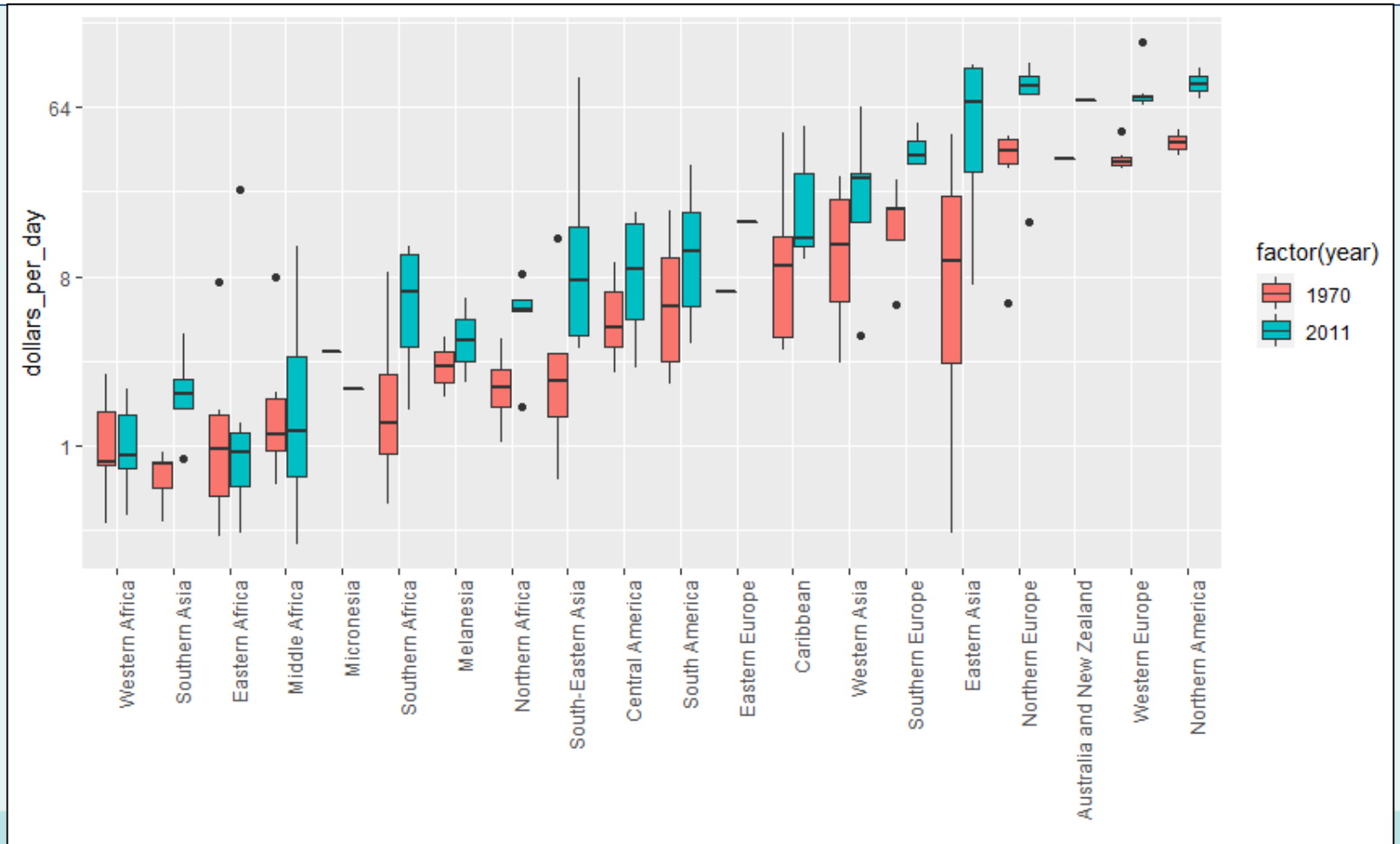
```
gapminder %>% filter(year %in% c(past_year, present_year) &
  (country %in% country_list) %>%
  mutate(group = ifelse(region %in% west, "Occidente", "Subdesarrollo")) %>%
  ggplot(aes(dollars_per_day)) +
  geom_histogram(binwidth = 1, color = "black") +
  scale_x_continuous(trans = "log2") +
  facet_grid(year ~ group)
```

¿Con el paso del tiempo los países ricos se hicieron más ricos y los pobres más pobres?



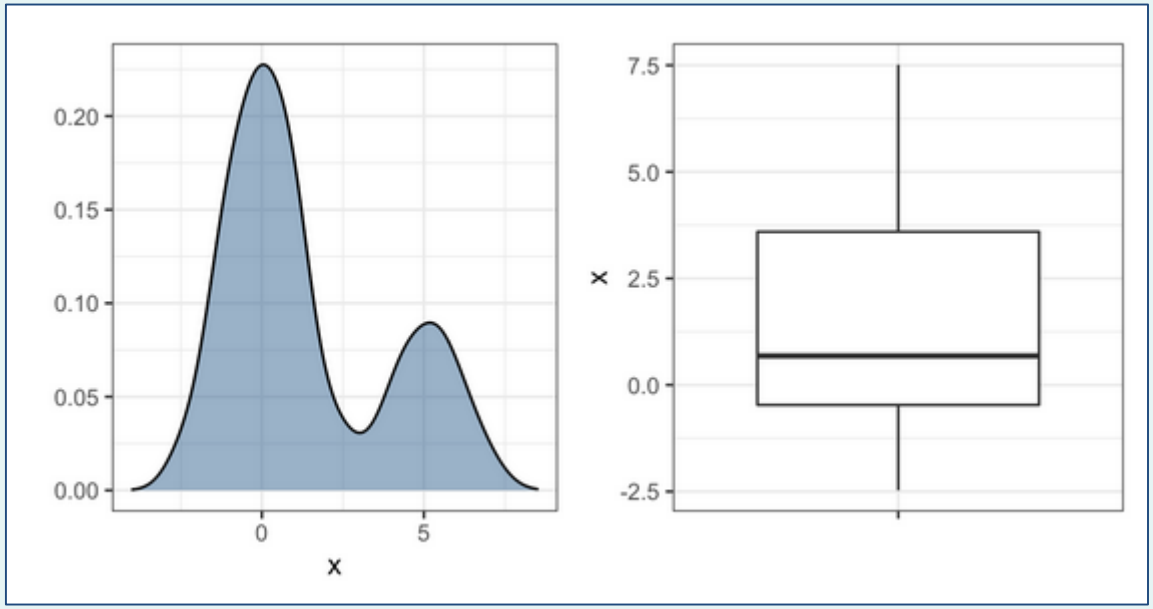
Otra comparación...

```
p <- gapminder %>% filter(year %in% c(past_year, present_year) & country %in%
country_list) %>% mutate(region = reorder(region, dollars_per_day, FUN =
median)) %>% ggplot() + theme(axis.text.x = element_text(angle = 90, hjust =
1)) + xlab("") + scale_y_continuous(trans = "log2")
p + geom_boxplot(aes(region, dollars_per_day, fill = factor(year)))
```



Elegir gráficos implica seleccionar visualización y datos

- Eligiendo un gráfico de cajas que es un resumen de los datos podemos perder características de la distribución de los mismos



- Distribución bimodal: En este caso perdemos los 2 modos de la distribución

Gráfico de densidad

- Transmitir el mensaje anterior a través de un gráfico de densidad
- Distribución de ingresos de 1970 vs 2011

```
gapminder %>% filter(year %in% c(past_year, present_year) &
                (country %in% country_list) )%>% ggplot(aes(dollars_per_day))
+
  geom_density(fill="black",alpha = 0.2 ) + scale_x_continuous(trans = "log2")
+ facet_grid(. ~ year)
```

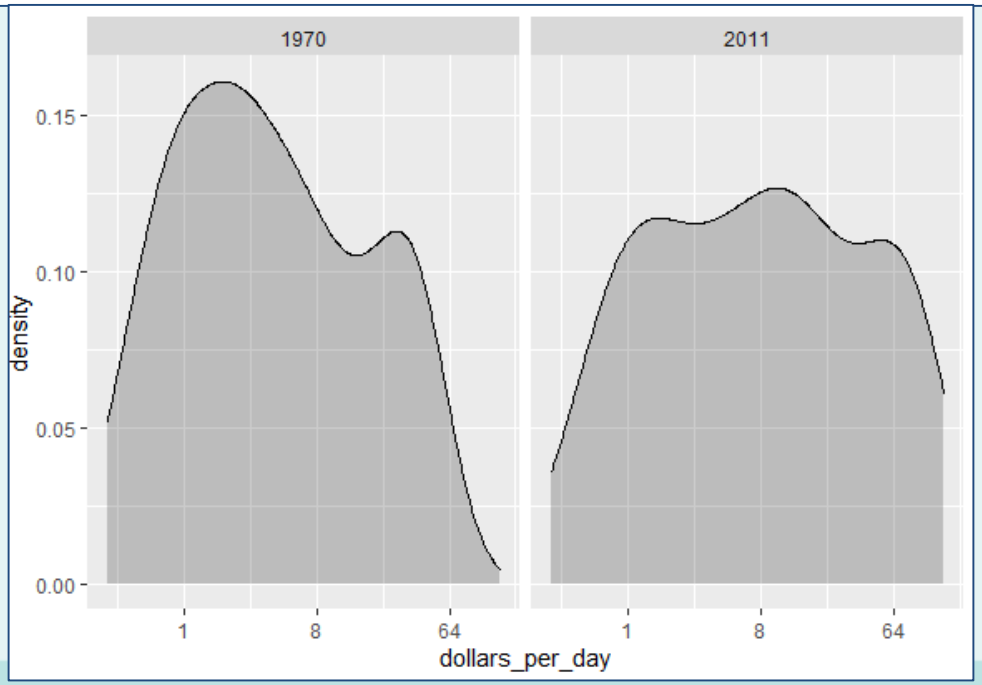
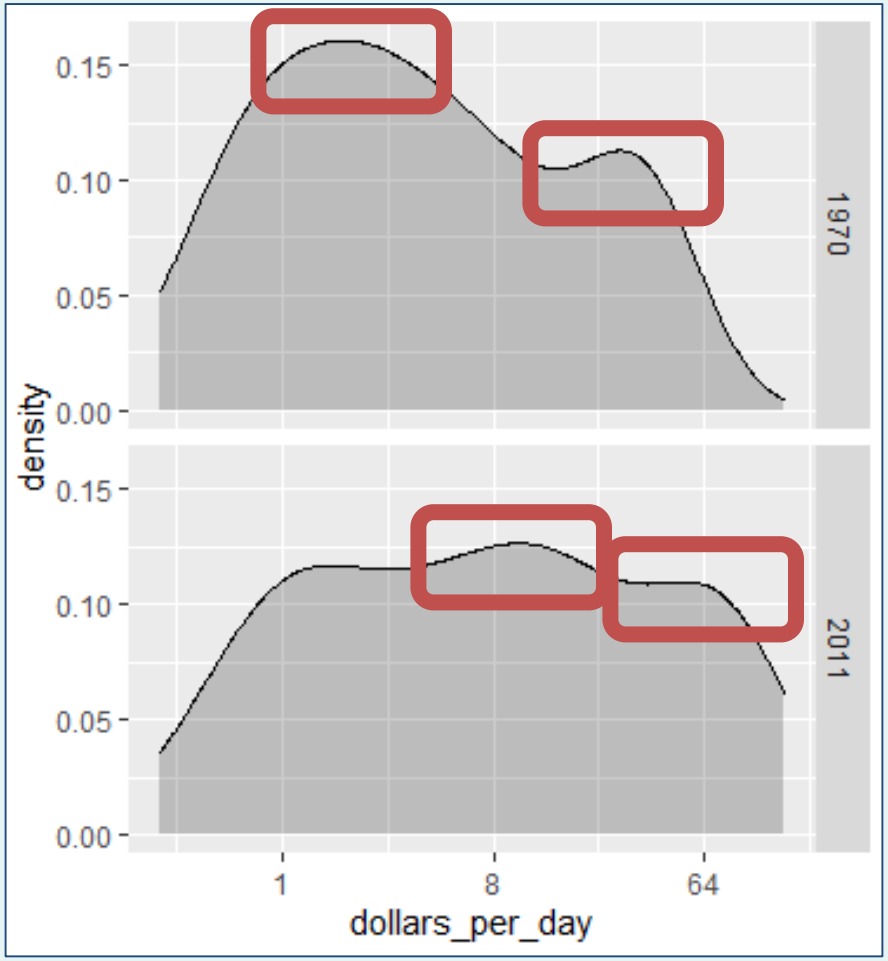


Gráfico de densidad - características en distribución de datos



Ponderar datos en gráficos de distribución

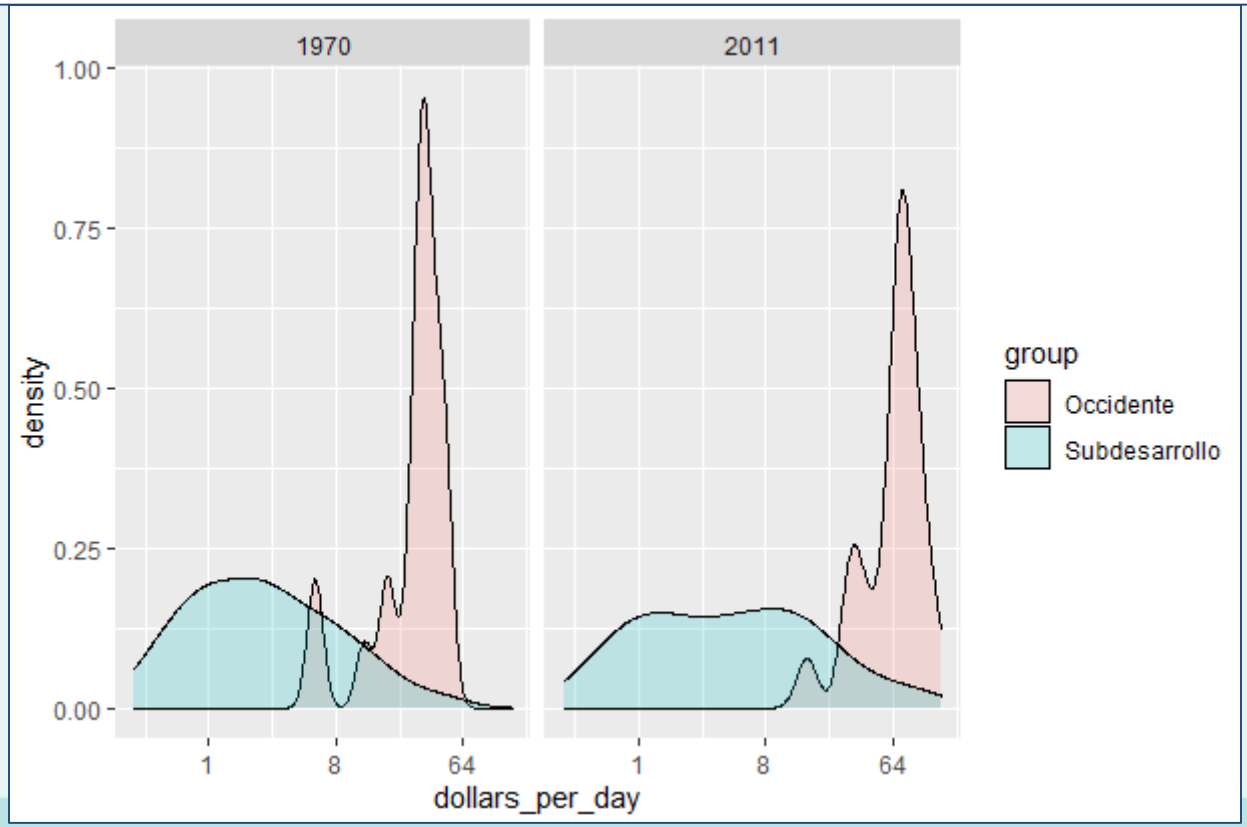
- A veces necesitamos que los gráficos de densidad preserven información de cuantos sujetos hay en cada grupo.
- Por defecto el área bajo la curva es 1 independientemente del tamaño de cada grupo
- En el ejemplo necesitamos saber cuantos paises hay en cada grupo

```
gapminder %>%
   filter(year == past_year & country %in% country_list) %>%
   mutate(group = ifelse(region %in% west,"Occidente", "Subdesarrollo")) %>%
group_by(group) %>%
   summarize(n = n()) %>% knitr::kable()
```

group	n
occidente	21
subdesarrollo	83

Gráficos de distribución sin ponderar

```
gapminder %>% filter(year %in% c(past_year, present_year) &(country %in%
   country_list) )%>% mutate(group = ifelse(region %in% west,
"Occidente",
   "Subdesarrollo")) %>% ggplot(aes(dollars_per_day,fill = group)) +
   geom_density(alpha = 0.2) + scale_x_continuous(trans = "log2")+
   facet_grid(. ~ year)
```



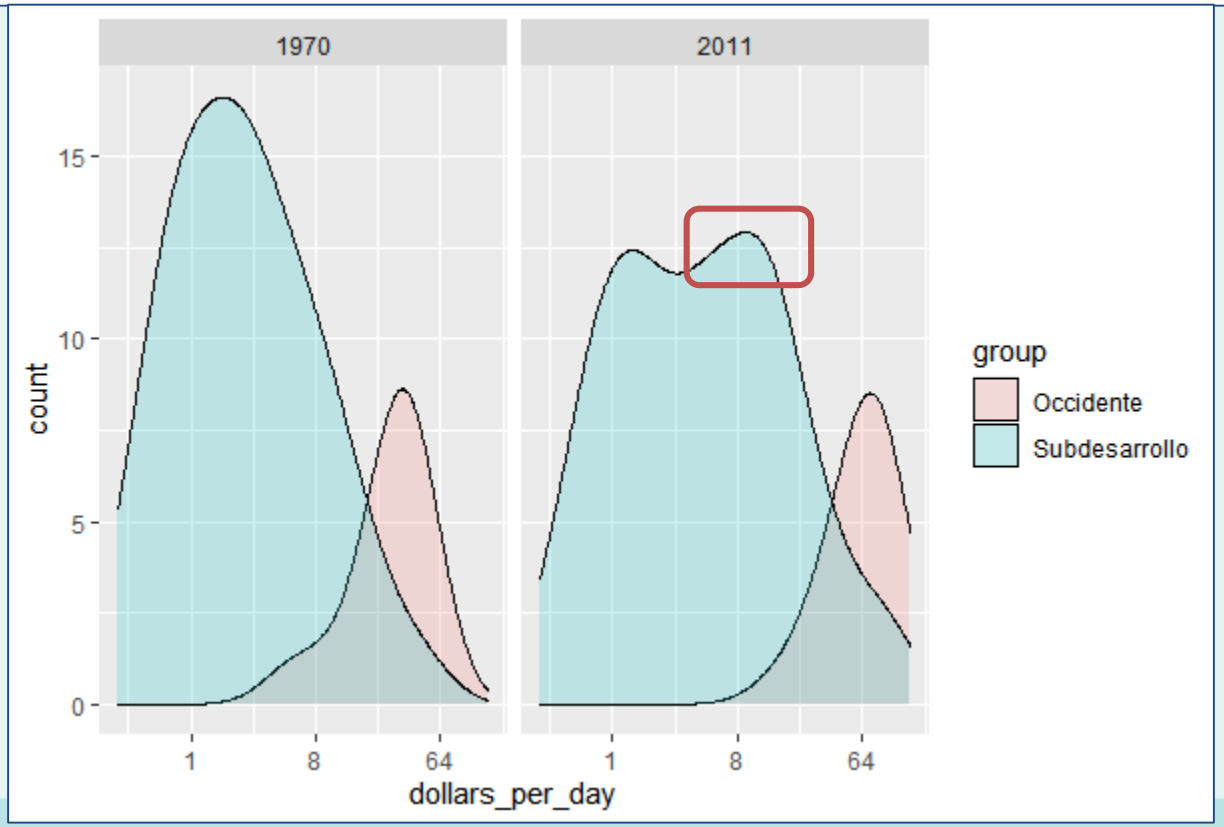
Cantidad de datos en gráficos

- Podemos cambiar el eje y del gráfico para que sea la cantidad de sujetos utilizando el operador '..'
- En el ejemplo si mapeamos

```
aes(x=dollars_per_day,y= ..count..)
```

Distribución ponderada y suavizada

```
gapminder %>% filter(year %in% c(past_year, present_year) &
 (country %in% country_list) )%>%
 mutate(group = ifelse(region %in% west,"Occidente","Subdesarrollo")) %>%
 ggplot(aes(dollars_per_day,fill = group, y = ..count..)) +
 geom_density(alpha = 0.2, bw = 0.75) +
 scale_x_continuous(trans = "log2") + facet_grid(. ~ year)
```



Función case_when

- Generar nueva información a través de datos y condiciones lógicas if - else if - else
- Nos permite vectorizar las funciones ‘if’ y ‘esle if’

```
case_when(condición1~salida1,
  condición2~salida2,....,
  condiciónN~salidaN,
  TRUE~salidaElse)
```

Condición y salida 1

Condición y salida 2

Condición y salida N

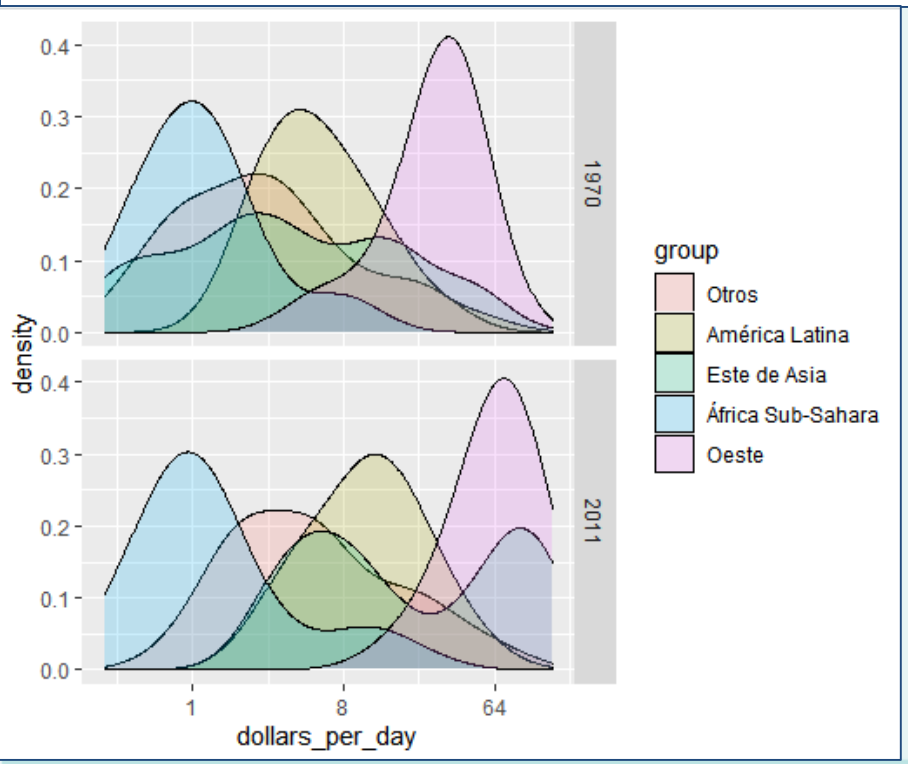
Salida si no se cumple ninguna de las condiciones anteriores

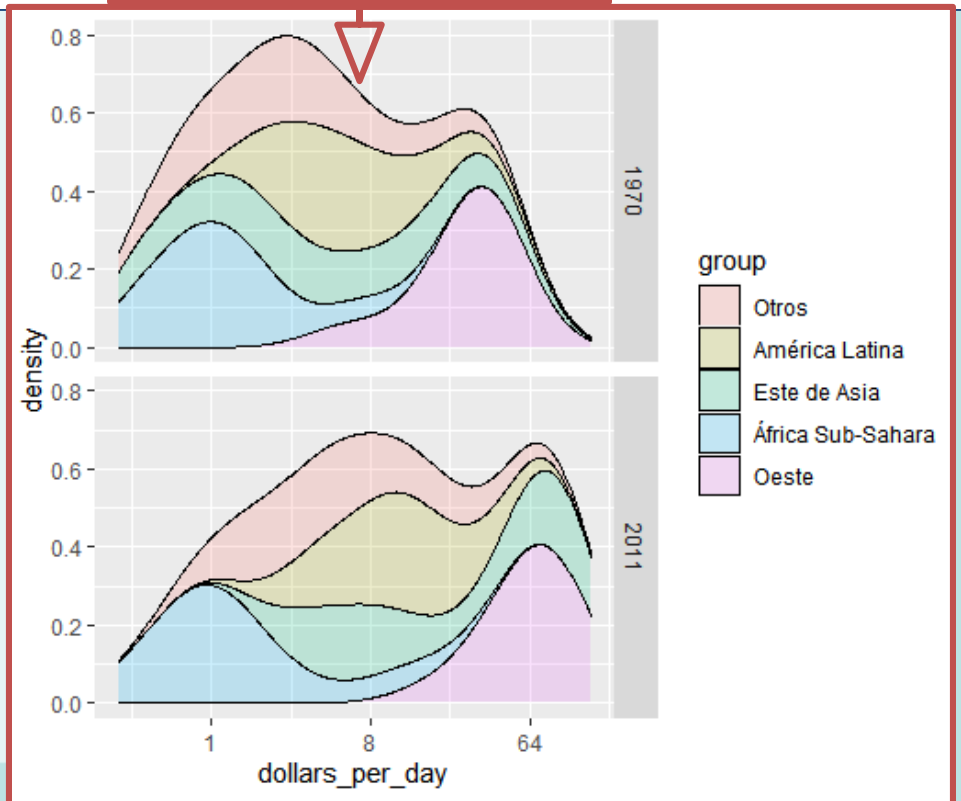
- Condición:
  - Operadores comparativos ej >=
  - Expresiones lógicas y(&), o(|), not(!)

Ejemplo case-when

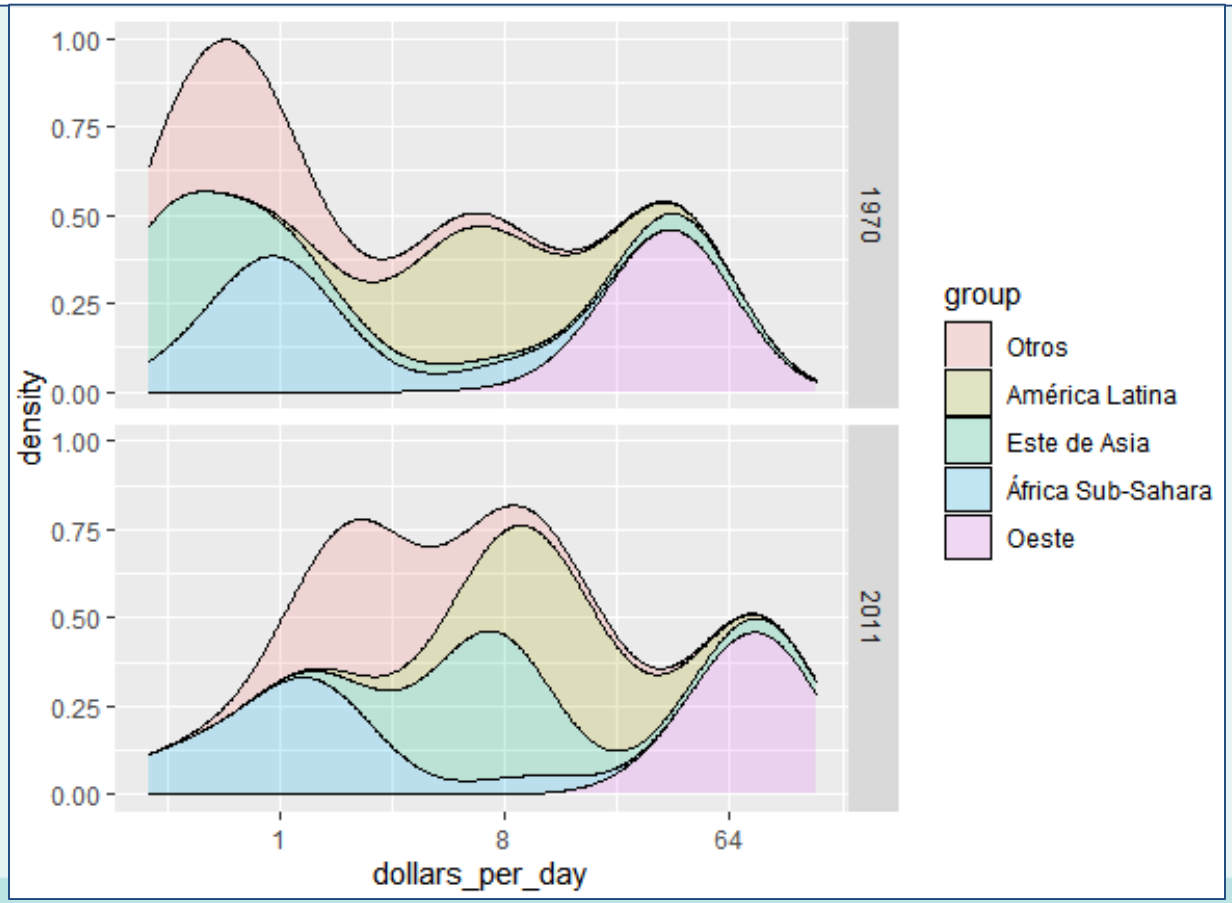
```
gapminder <- gapminder %>%
  mutate(group = case_when(
 .$region %in% west ~ "Oeste",
.$region %in% c("Eastern Asia","South-Eastern Asia") ~
"Este de Asia",
      .$region %in% c("Caribbean", "Central America",
"South America") ~ "América Latina",
      .$continent == "Africa" & .$region != "Northern
Africa" ~ "África Sub-Sahara",
      TRUE ~ "Otros"))
```

```r
# re ordenamos - niveles del factor
gapminder <- gapminder %>%
  mutate(group = factor(group, levels = c("Otros", "América Latina",
"Este de Asia", "África Sub-Sahara", "Oeste")))
# Re definimos el objeto
p <- gapminder %>% filter(year %in% c(past_year, present_year) & country
%in% country_list) %>% ggplot(aes(dollars_per_day, fill = group)) +
scale_x_continuous(trans = "log2")
p + geom_density(alpha = 0.2, bw = 0.75, position = "stack") +
  facet_grid(year ~ .)
```





```
gapminder %>% filter(year %in% c(past_year, present_year) &
country %in% country_list) %>% group_by(year) %>% mutate(weight =
population/sum(population*2)) %>% ungroup() %>%
ggplot(aes(dollars_per_day, fill = group,weight = weight)) +
  scale_x_continuous(trans = "log2") +
  geom_density(alpha = 0.2, bw = 0.75, position = "stack") +
facet_grid(year ~ .)
```



Comunicar erróneamente datos

- Muchas veces mostrar el promedio de datos es extraer conclusiones erróneas para los miembros involucrados.


infant_survival_rate
0.998
0.995
0.990
0.950
0.900
0.850
1
The West
Latin America
Northern Africa
Southern Asia
Sub-Saharan Africa
1 - infant_mortality/1000
0.998
0.995
0.990
0.950
0.900
0.850
1
8
64
dollars_per_day
Singapore
Sweden
Serbia
United States
Chile
Mauritius
Tunisia
Cambodia
Bolivia
Botswana
Sudan
Haiti
Sierra Leone
Angola
group
East Asia
Latin America
Northern Africa
Pacific Islands
Southern Asia
Sub-Saharan Africa
The West