Tarea individual 1

Entregar el Lunes 16 de Abril

9/4/2018

Entrega

La tarea debe ser realizada en RMarkdown en un repositorio de GitHub llamado "Tarea 1". La tarea es individual por lo que cada uno tiene que escribir su propia versión de la tarea. El repositorio debe contener unicamente el archivo .Rmd con la solución de la tarea. Vamos a utilizar la librería `gapminder`, por lo que si no la usaste anteriormente tenés que instalarla y luego cargarla. Para obtener la descripción del paquete `library(help = "gapminder")` y para saber sobre la base `?gapminder`.

Idea básica de regresión lineal

Una regresión lineal es una aproximación utilizada para modelar la relación entre dos variables que llamaremos `X` e `Y`. Donde Y es la variable que queremos explicar y X la variable explicativa (regresión simple).

El análisis de regresión ajusta una curva a travéz de los datos que representa la media de Y dado un valor especificado de X. Si ajustamos una regresión lineal a los datos consideramos "la curva media" como aquella que mejor ajusta a los datos.

Algunas veces ajustamos curvas genéricas promediando puntos cercanos entre si con métodos de suavizado no necesariamente lineales. ¿Cómo incluimos una recta de regresión en nuestro gráfico?

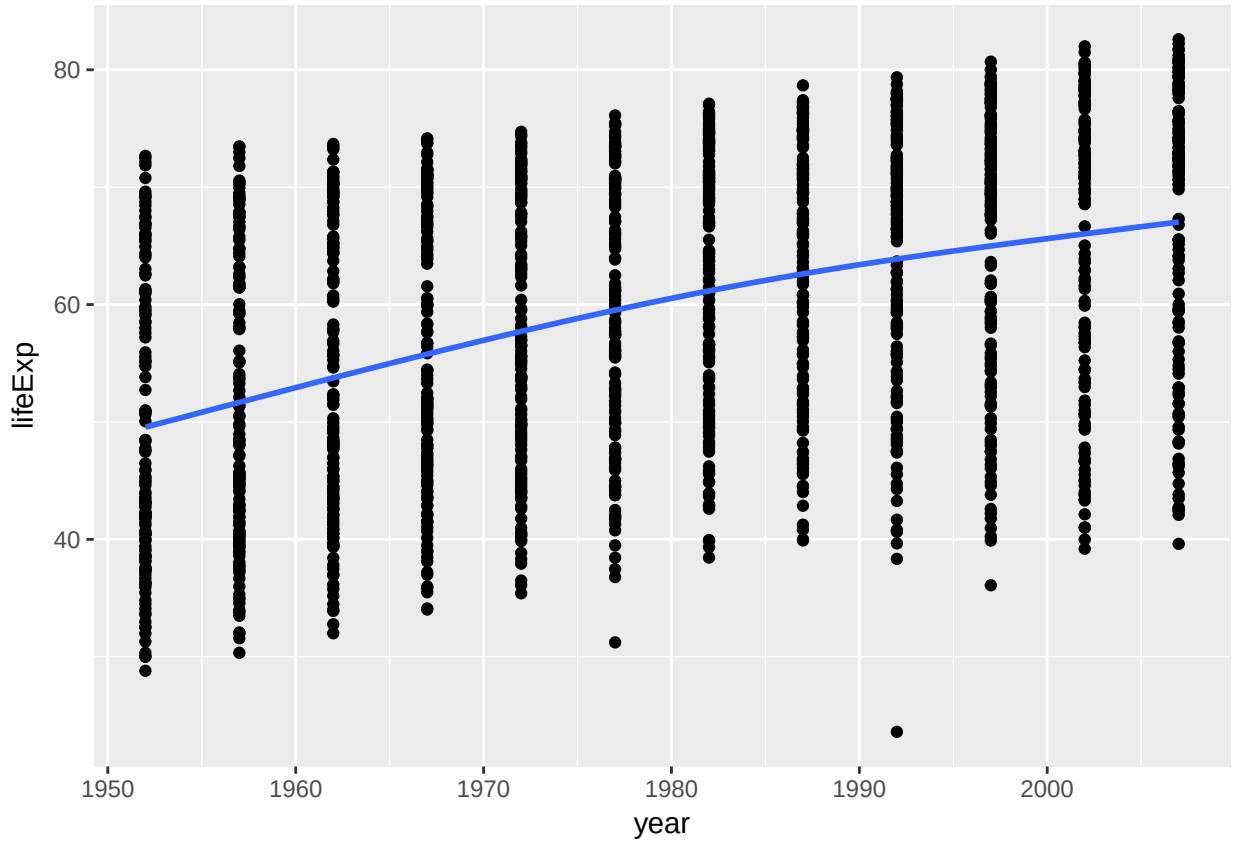
ajustamos una recta de regresión a los datos en Para agregar una linea de regresión o una curva tinenes que agregar una capa a tu gráfico `geom_smoth`. Probablemente dos de los argumentos más útiles de `geom_smoth` son:

- `method = ...`
  - ...“lm” para una linea recta. `lm` “Linear Model”.
  - ...otro para una curva genérica (llamada de suavizado; por defecto, es la parte `smooth` de `geom_smooth`).
  - `se=...` controla si los intervalos de confianza son dibujados o no.

Ejemplo:

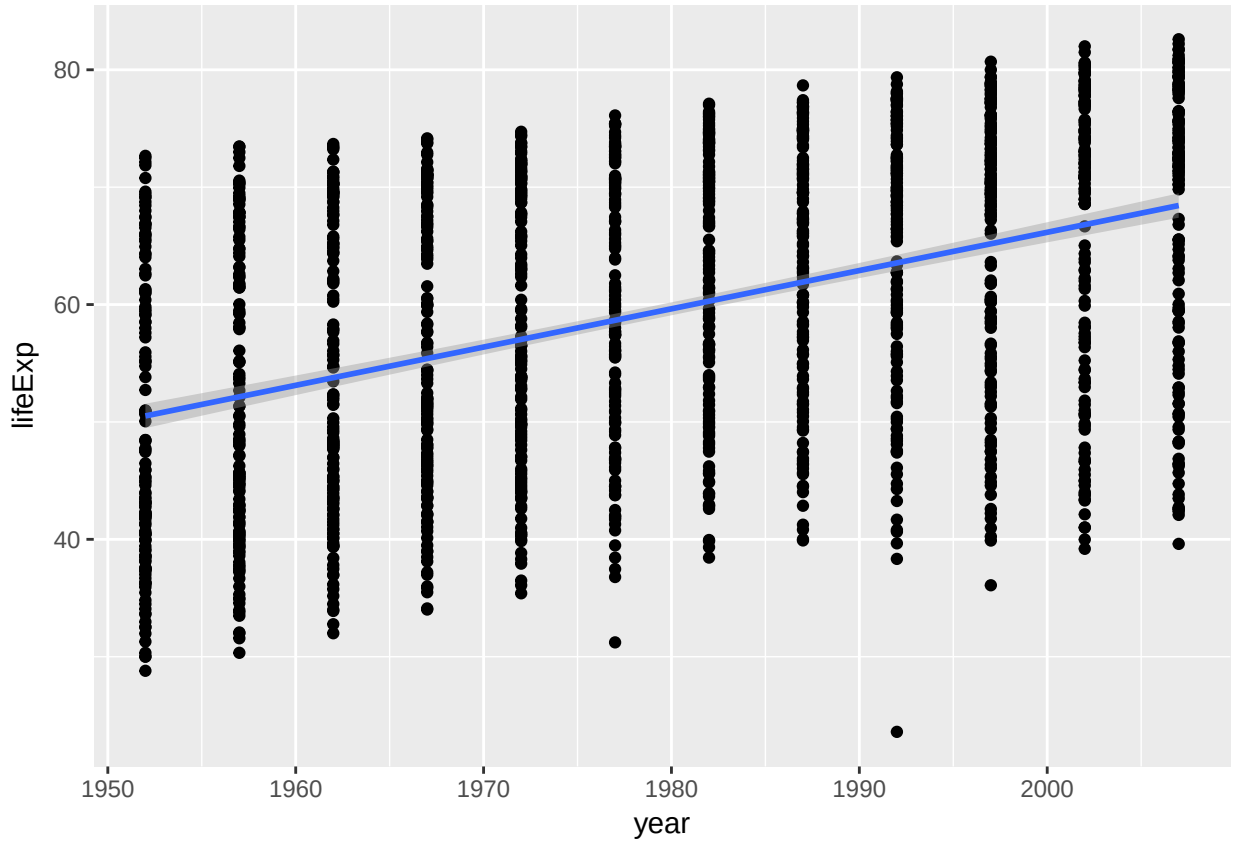
```
vc1 <- ggplot(gapminder, aes(year, lifeExp)) +
    geom_point()
vc1 + geom_smooth(se = FALSE)
```

```
## `geom_smooth()` using method = 'gam'
```



En este caso `geom_smooth()` está usando `method = 'gam'`

```
vc1 + geom_smooth(method = "lm")
```



Ejercicio 1

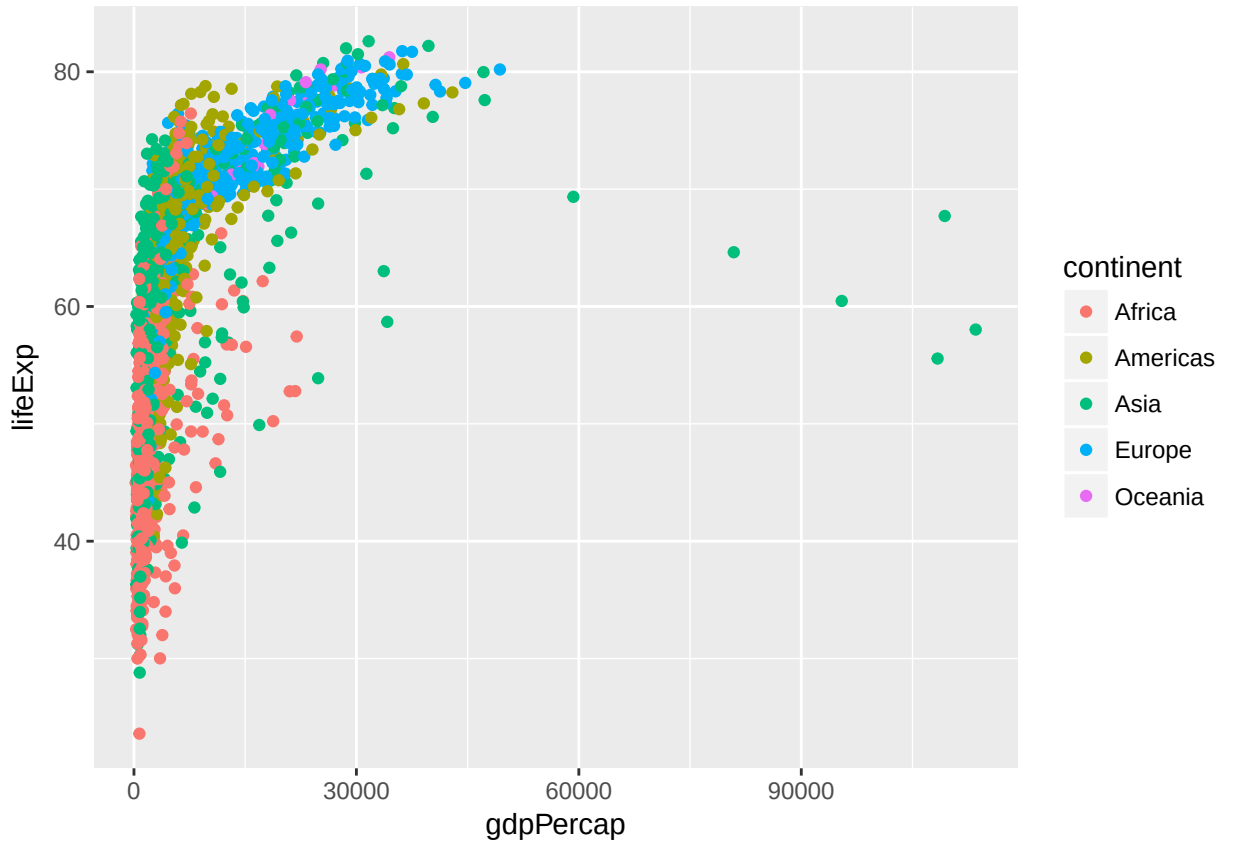
Hacer un gráfico de dispersión que tenga en el eje `y year` y en el eje `x lifeExp`, los puntos deben estar coloreados por la variable `continent`. Para este plot ajustá una recta de regresión para cada continente sin incluir las barras de error. Las etiquetas de los ejes deben ser claras y describir las variables involucradas. Incluir un `caption` en la figura con algún comentario de interés que describa el gráfico.

Ejercicio 2

Omitir la capa de `geom_point` del gráfico anterior. Las lineas aún aparecen aunque los puntos no. ¿Porqué sucede esto?

Ejercicio 3

El siguiente es un gráfico de dispersión entre `lifeExp` y `gdpPercap` coloreado por la variable `continent`. Usando como elemento estético color (`aes`) nosotros podemos distinguir los distintos continentes usando diferentes colores de similar manera usando forma (`shape`).



El gráfico anterior está sobrecargado, ¿de que forma modificarías el gráfico para que sea más clara la comparación para los distintos continentes y porqué? Las etiquetas de los ejes deben ser claras y describir las variables involucradas. Comentá alguna característica interesante que describa lo que aprendes viendo el gráfico.

Ejercicio 4

Hacer un gráfico de lineas que tenga en el eje `x year` y en el eje `y gdpPercap` para cada continente en una misma ventana gráfica. En cada continente, el gráfico debe contener una linea para cada país a lo largo del tiempo (serie de tiempo de `gdpPercap`). Las etiquetas de los ejes deben ser claras y describir las variables involucradas. Incluir un `caption` en la figura con algún comentario de interés que describa el gráfico.

Ejercicio 5

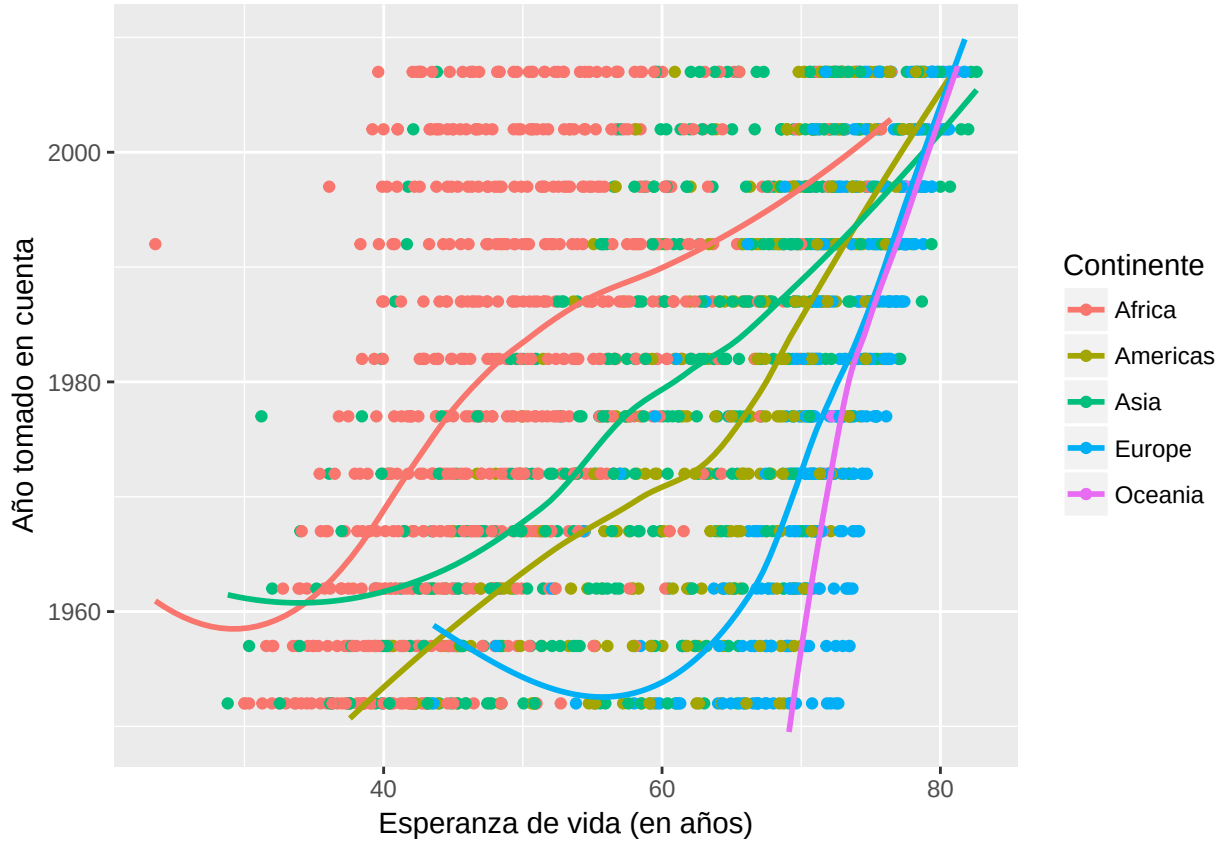
Usando los datos `gapminder` seleccione una visualización que describa algún aspecto de los datos que no exploramos. Comente algo interesante que se puede aprender de su gráfico.

Estudiante: Emiliano Barone

Ejercicio 1

```
ggplot(gapminder, aes(lifeExp, year, colour=continent))+
  geom_point() +
  geom_smooth(se=F) +
  labs(x="Esperanza de vida (en años)" ,y ="Año tomado en cuenta", colour="Continente")
```

```
## `geom_smooth()` using method = 'loess'
```



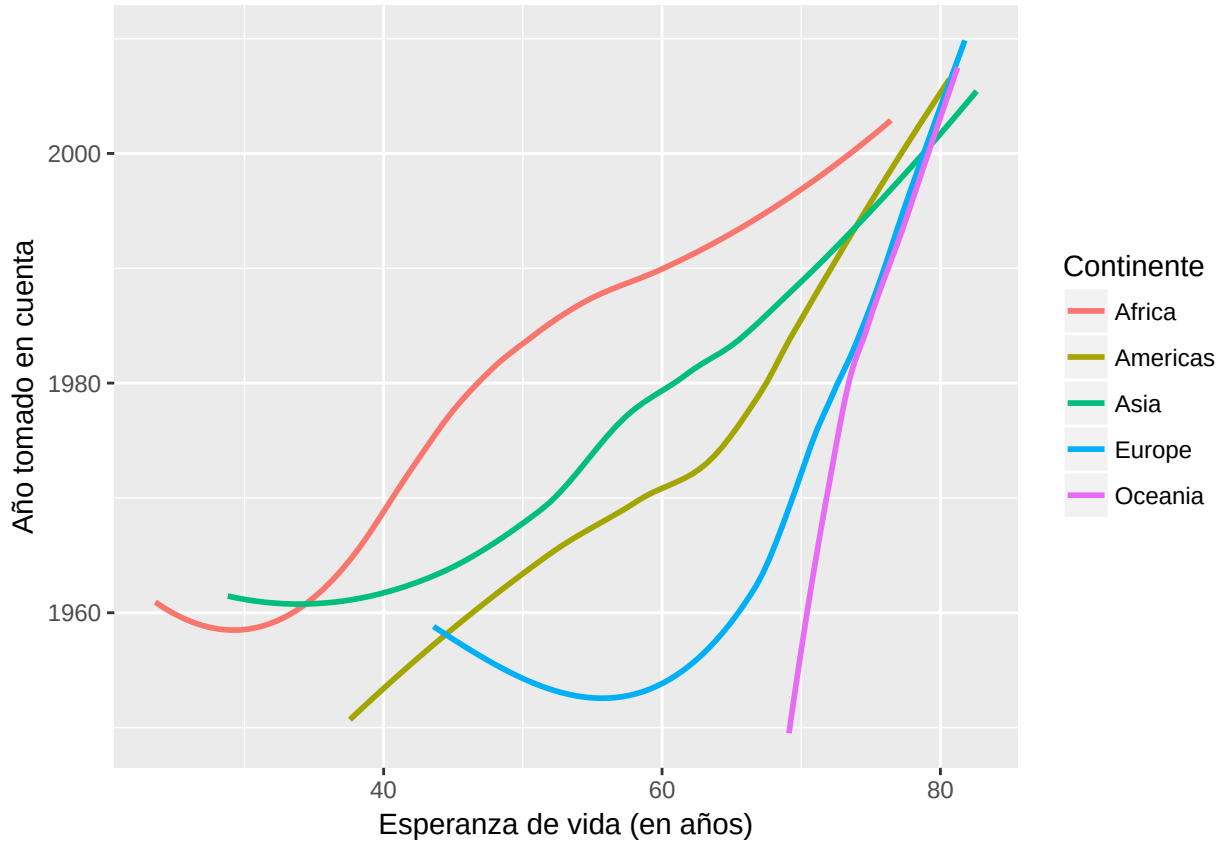
En este grafico se puede apreciar como varia la esperanza de vida a lo largo del tiempo dependiendo del continente bajo estudio. El caso que mas resalta es el del continente de Oceania, el cual siempre mantuvo una muy buena esperanza de vida, aunque en los alrededores del año 2005 se vio levemente superada por Asia y posteriormente por Europa.

El análisis de regresión ajusta una curva a travéz de los datos que representa la media de Y dado un valor especificado de X. Si ajustamos una regresión lineal a los datos consideramos "la curva media" como aquella que mejor ajusta a los datos.

Ejercicio 2

```
ggplot(gapminder, aes(lifeExp, year, colour=continent))+
  geom_smooth(se=F) +
  labs(x="Esperanza de vida (en años)" ,y ="Año tomado en cuenta", colour="Continente")
```

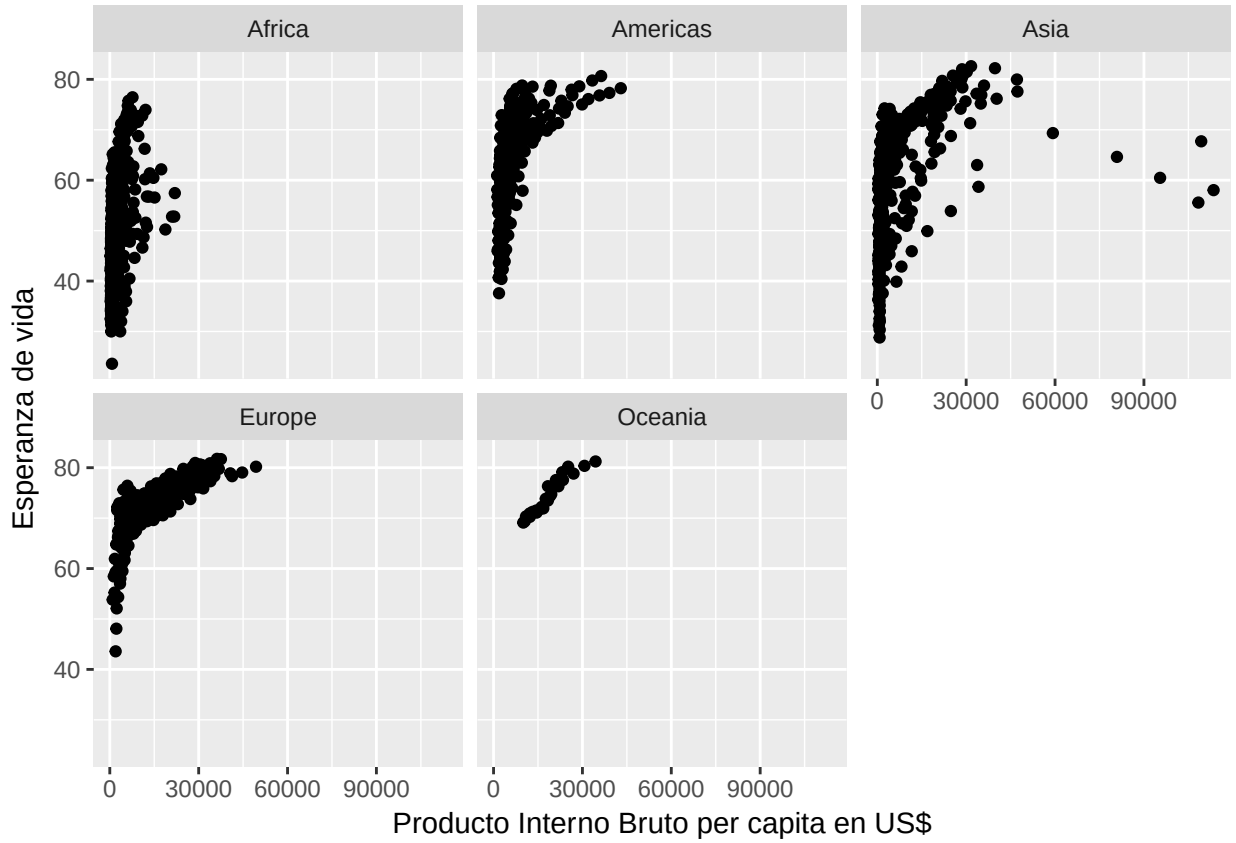
```
## `geom_smooth()` using method = 'loess'
```



El hecho de que aunque se haya quitado la capa del **geom_point()** se vean las curvas de regresion es debido a que en las lineas de codigo ya se le indico al programa la base de datos con la cual se quiere trabajar y las variables a analizar. Como el programa tiene esta informacion puede calcular sin problema los distintos valores promedio de una variable para un valor fijo de la otra. Al utilizar la funcion **geom_smooth()** le estamos pidiendo que grafique una curva que sera la que mejor se ajuste a los datos, y como resultado obtenemos unicamente las lineas sin los puntos ya que omitimos la capa de la gramatica grafica que los mostraba.

Ejercicio 3

```
ggplot(gapminder, aes(gdpPercap, lifeExp))+
  geom_jitter()+
  facet_wrap(~continent)+
  labs(x="Producto Interno Bruto per capita en US$" ,y ="Esperanza de vida", colour="Continente")
```

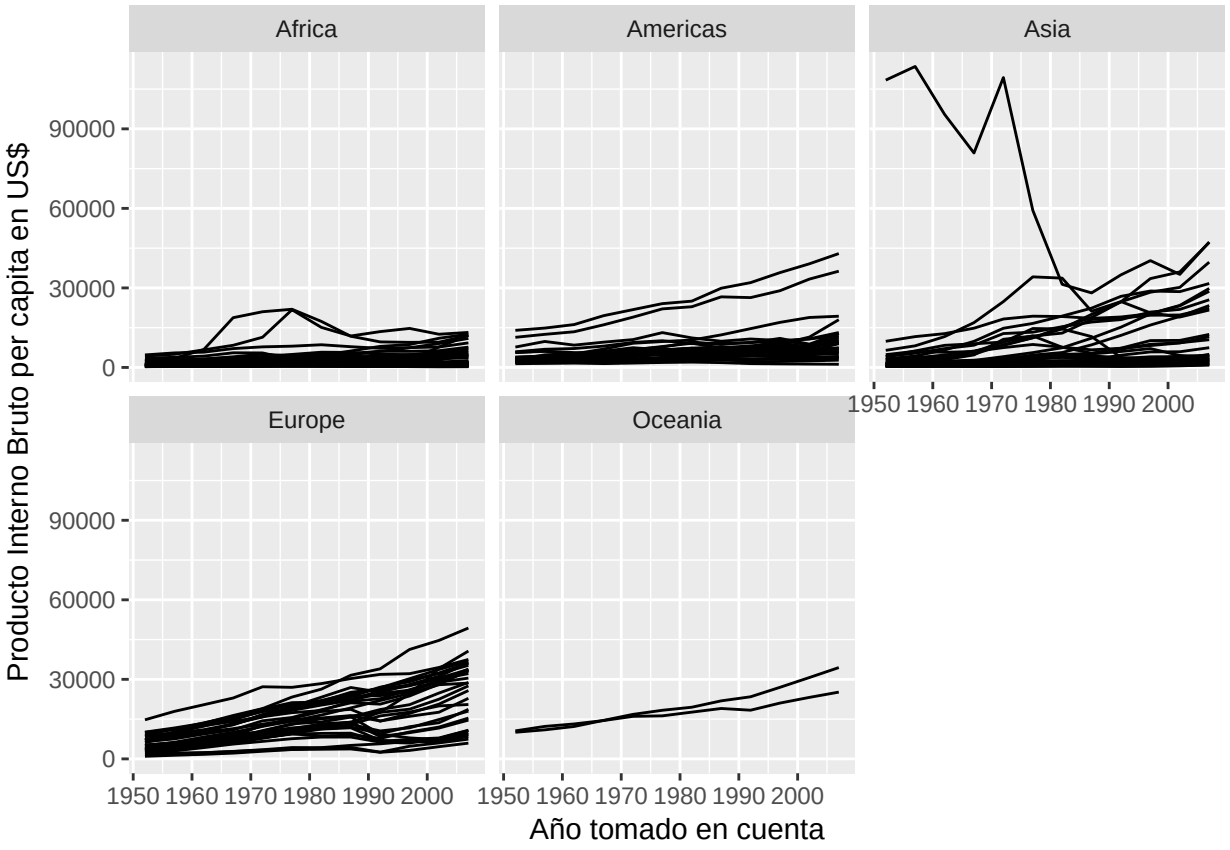


Segun lo escrito en su libro **"Elegant graphics for data analysis"** el experto en datos **Hadley Wickham** menciona que hay tres maneras de lidiar con el sobreploteo. Una es utilizando la funcion **geom_jitter()**, la otra es utilizando los ya conocidos boxplot(diagrama de cajas),y finalmente la ultima es utilizando el vioplot(grafico del violin). Para este caso especifico es conveniente utilizar la funcion **geom_jitter()** ya que es la mas representativa y tambien la mas entendible. El grafico de cajas solamente nos resume la distribucion de los datos con cinco numeros, y el grafico de violin es de comprension compleja.

Tambien decidimos separar el grafico en subcategorias, ya que una de las desventajas de la funcion **geom_jitter()** es que no sirve para cantidades muy grandes de datos. Al separar por continente con la funcion **facet_wrap()** esta desventaja se ve disminuida y es mas facil hacer comparaciones entre continentes. La conclusion mas importante de este grafico es que en la mayoria de los continentes parece haber una relacion entre el PBI per capita y la esperanza de vida, es decir, a mayor esperanza de vida mayor PBI per capita. El unico continente en el cual no se nota claramente esta relacion es el continente africano.

Ejercicio 4

```
ggplot(gapminder , aes(x = year, y = gdpPercap, group=country)) +
  geom_line() + facet_wrap(~ continent) +
  labs(x="Año tomado en cuenta" ,y ="Producto Interno Bruto per capita en US$")
```



En el siguiente grafico podemos ver la evolucion del PBI per capita por pais en cada continente a lo largo del tiempo. Podemos notar que el continente europeo es en el cual el Pbi per capita parece aumentar cada vez mas a lo largo del tiempo. En el continente americano hay un especie de comportamiento general y dos casos que presentan mayor crecimiento, los cuales parecerian ser los paises de Estados Unidos y Canada. En el caso del continente asiatico, a principios de los años 1950 un pais presenta un pbi considerado extremadamente alto en comparacion con los otros paises del mundo.

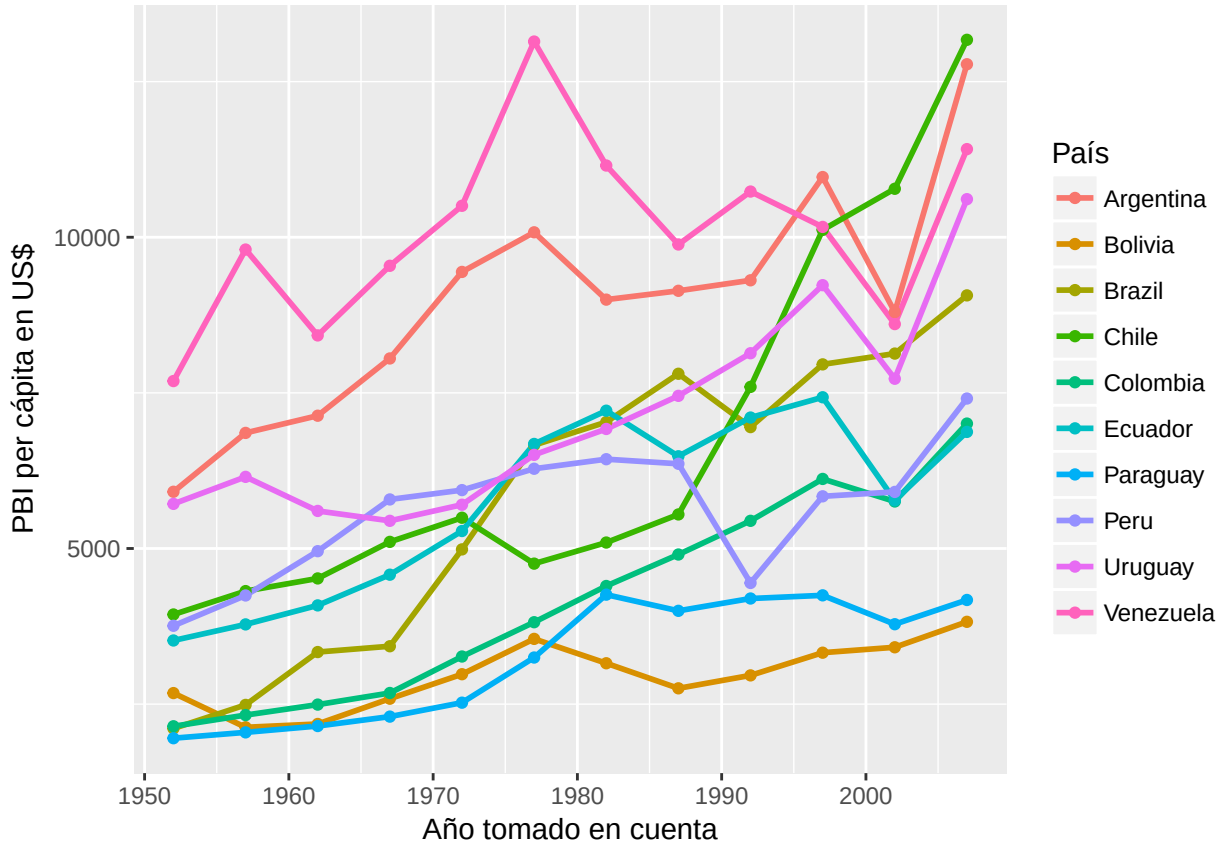
```
dplyr::filter(gapminder, year<1990, gdpPercap>80000)
```

```
## # A tibble: 5 x 6
##   country continent  year lifeExp    pop gdpPercap
##   <fct>   <fct>     <int>   <dbl>  <int>     <dbl>
## 1 Kuwait  Asia       1952    55.6 160000   108382.
## 2 Kuwait  Asia       1957    58.0 212846   113523.
## 3 Kuwait  Asia       1962    60.5 358266    95458.
## 4 Kuwait  Asia       1967    64.6 575003    80895.
## 5 Kuwait  Asia       1972    67.7 841934   109348.
```

Aqui podemos ver que el pais del continetne asiatico que presenta ese comportamiento es Kuwait el cual alcanzo un valor maximo de PBI per capita de US$113.523,13 en el año 1957.

Ejercicio 5

```
ggplot(AmericaS, aes(year, gdpPercap, colour=country))+
  geom_line(size=1)+
  geom_point()+
  labs(x="Año tomado en cuenta" , y ="PBI per cápita en US$", colour="País")
```



En este grafico podemos analizar como varia el PBI per capita a lo largo de los años en los distintos paises de América del sur. Es interesante denotar que ningun pais ha tenido un crecimiento constante del PBI. Los casos de Bolivia y Paraguay se han mantenido ultimos durante varios años. En el caso de nuestro pais (Uruguay), se nota que aproximadamente desde el año 1960 su PBI ha crecido durante años hasta al año 2002, esto probablemente sea debido a la crisis ya que años despues se puede ver como continua creciendo.