Tarea 2

Ángela Vieyto 5.487.839-8

Entrega 26 de Abril

Ejercicio 1

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. El resto de los comentarios del gráfico se realizan en el texto.***

```
gapminder %>%
    ggplot(aes(x = lifeExp, y = year, color = continent)) +
    geom_point(size = 1, alpha = 0.5) + geom_smooth(method = "lm",
    se = FALSE) + labs(x = "Esperanza de vida al nacer",
    y = "Año", color = "Continente", title = "G1. Evolución de la esperanza de vida al nacer",
    subtitle = "Período 1952 - 2007", caption = "Relación entre la esperanza de vida al nacer y el tiem
    scale_color_discrete(labels = c(Africa = "África",
        Americas = "América", Europe = "Europa", Oceania = "Oceanía")) +
    ylim(1950, 2010) + theme(plot.title = element_text(face = "bold",
    size = 15), axis.title = element_text(face = "bold"),
    legend.title = element_text(face = "bold"))
```

Relación entre la esperanza de vida al nacer y el tiempo, distinguiendo según continente.

Observando el gráfico G1 encontramos una relación positiva entre la esperanza de vida al nacer y el tiempo. Es decir, con el transcurso de los años la esperanza de vida promedio ha ido en aumento.

Analizando la pendiente de las rectas de regresión podríamos decir que el incremento en la esperanza de vida promedio ha sido más significativo en África y Asia en comparación con Oceanía y Europa.

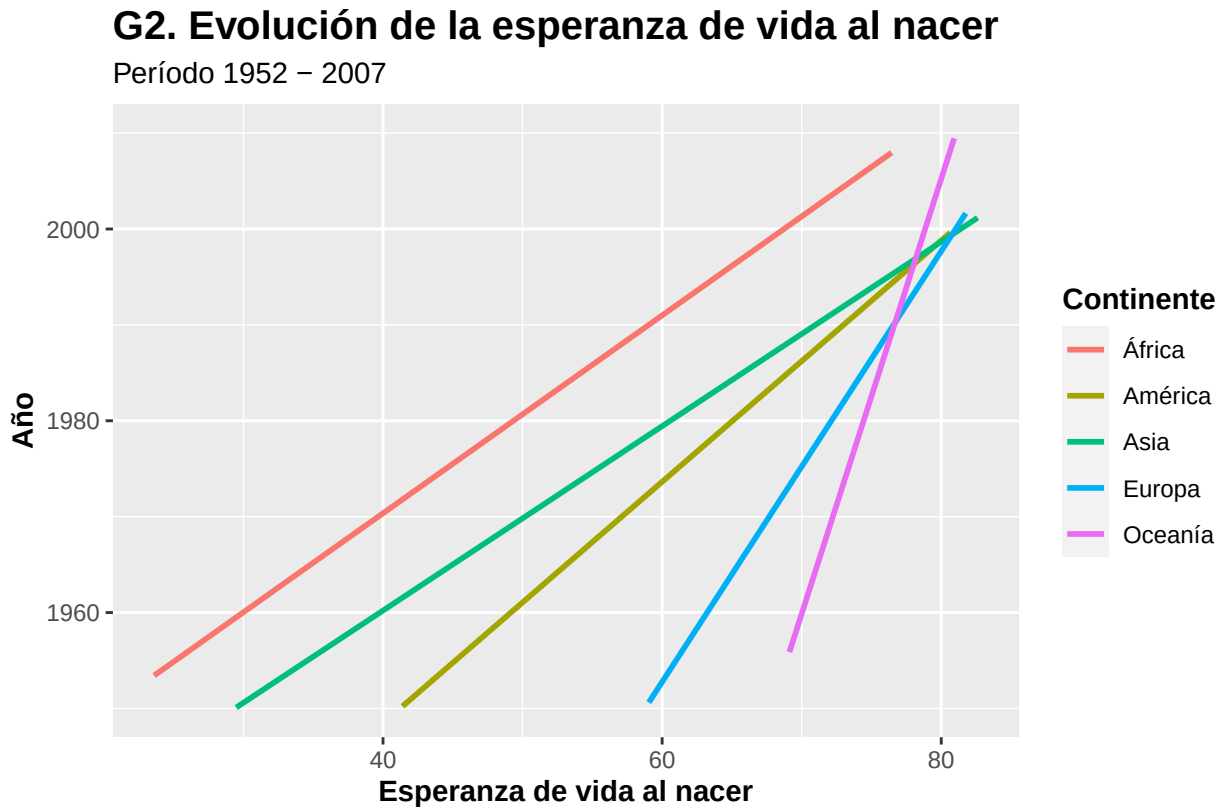
Es interesante notar que el primer grupo de continentes parte de valores significativamente menores que los del segundo grupo, lo cual intuitivamente parecería ser razonable. Podemos ver este aspecto en mayor detalle en la siguiente tabla:

```
gapminder %>%
    group_by(continent) %>%
    summarise(min_obs = min(lifeExp)) %>%
    arrange(min_obs)
```

```
## # A tibble: 5 x 2
##   continent min_obs
##   <fct>       <dbl>
## 1 Africa       23.6
## 2 Asia         28.8
## 3 Americas     37.6
## 4 Europe       43.6
## 5 Oceania      69.1
```

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

```r
gapminder %>%
    ggplot(aes(x = lifeExp, y = year, color = continent)) +
    geom_smooth(method = "lm", se = FALSE) + labs(x = "Esperanza de vida al nacer",
    y = "Año", color = "Continente", title = "G2. Evolución de la esperanza de vida al nacer",
    subtitle = "Período 1952 - 2007", caption = "Relación entre la esperanza de vida al nacer y el tiem
    scale_color_discrete(labels = c(Africa = "África",
        Americas = "América", Europe = "Europa", Oceania = "Oceanía")) +
    ylim(1950, 2010) + theme(plot.title = element_text(face = "bold",
    size = 15), axis.title = element_text(face = "bold"),
    legend.title = element_text(face = "bold"))
```

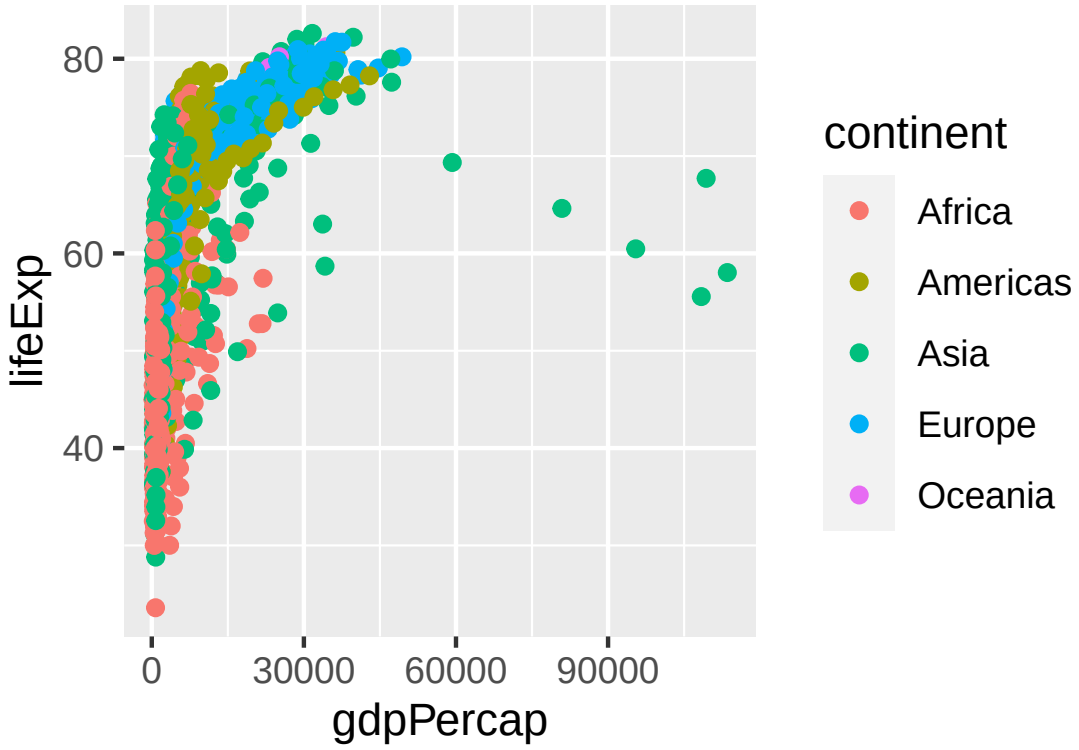


Relación entre la esperanza de vida al nacer y el tiempo, distinguiendo según continente.

Las líneas aparecen porque estamos usando la capa de `geom_smooth()`.

Las rectas de regresión graficadas por `geom_smooth()` y los puntos graficados por `geom_point()` son construidos a partir de los mismos datos, pero ambas capas se construyen de manera independiente. Es por este preciso motivo que una no depende de la otra y podemos utilizarlas por separado.

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`).***

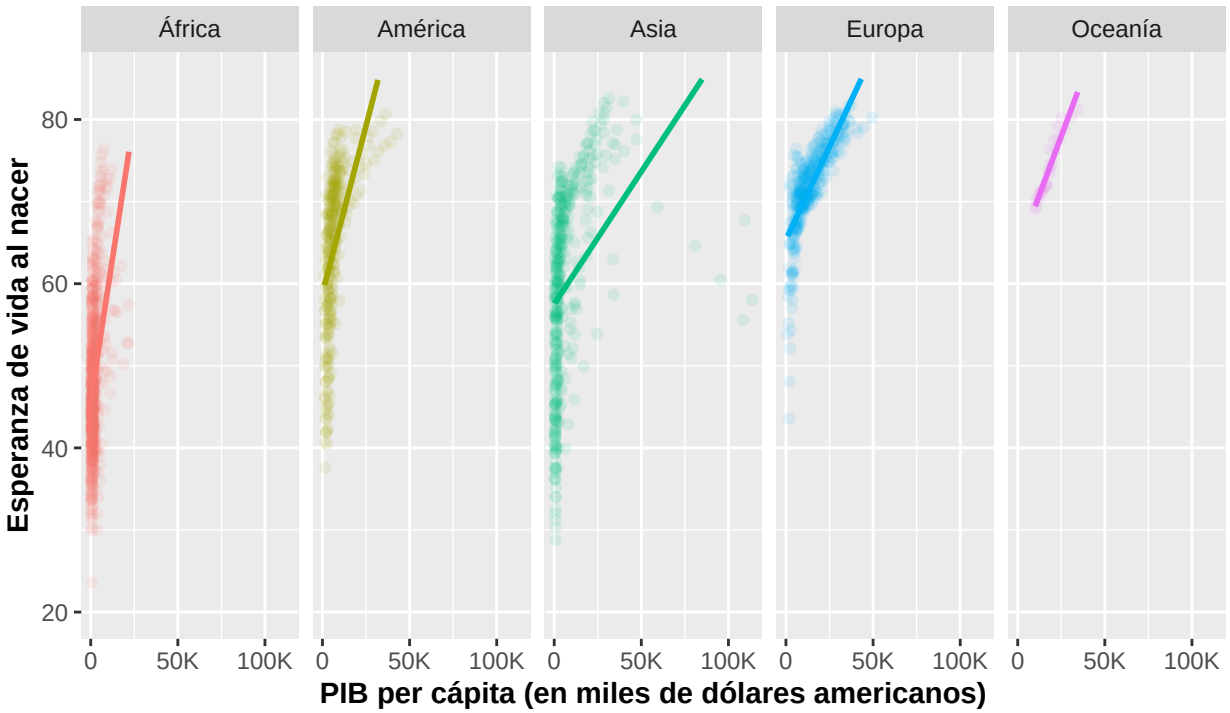

80
60
40
lifeExp
0
30000
60000
90000
gdpPercap
continent
Africa
Americas
Asia
Europe
Oceania

El gráfico anterior está sobrecargado, ¿de que forma modificarías el gráfico para que sea más clara la comparación para los distintos continentes y por qué? 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.

El sobreploteo existente dificulta que los continentes puedan distinguirse por color en el gráfico actual. En este sentido, la primer modificación sería aplicar una capa de `facet_grid()` para facilitar la visualización por continente.

También entiendo conveniente ajustar una recta de regresión lineal que permita visualizar rápidamente la relación entre ambas variables.

```
gapminder %>%
    ggplot(aes(x = gdpPercap, y = lifeExp, color = continent)) +
    geom_point(alpha = 0.1) + geom_smooth(method = "lm",
    se = FALSE) + labs(x = "PIB per cápita (en miles de dólares americanos)",
    y = "Esperanza de vida al nacer", color = "Continente",
    title = "G3. Esperanza de vida al nacer vs. PIB per cápita",
    subtitle = "Período 1952 - 2007", caption = "Relación entre la esperanza de vida al nacer y el PIB p
    theme(plot.title = element_text(face = "bold",
        size = 15), axis.title = element_text(face = "bold"),
        legend.title = element_text(face = "bold"),
        legend.position = "none") + facet_grid(cols = vars(continent),
    labeller = as_labeller(c(Africa = "África", Americas = "América",
        Asia = "Asia", Europe = "Europa", Oceania = "Oceanía"))) +
    scale_x_continuous(breaks = c(0, 50000, 1e+05),
        labels = c("0", "50K", "100K")) + ylim(20,
    85)
```



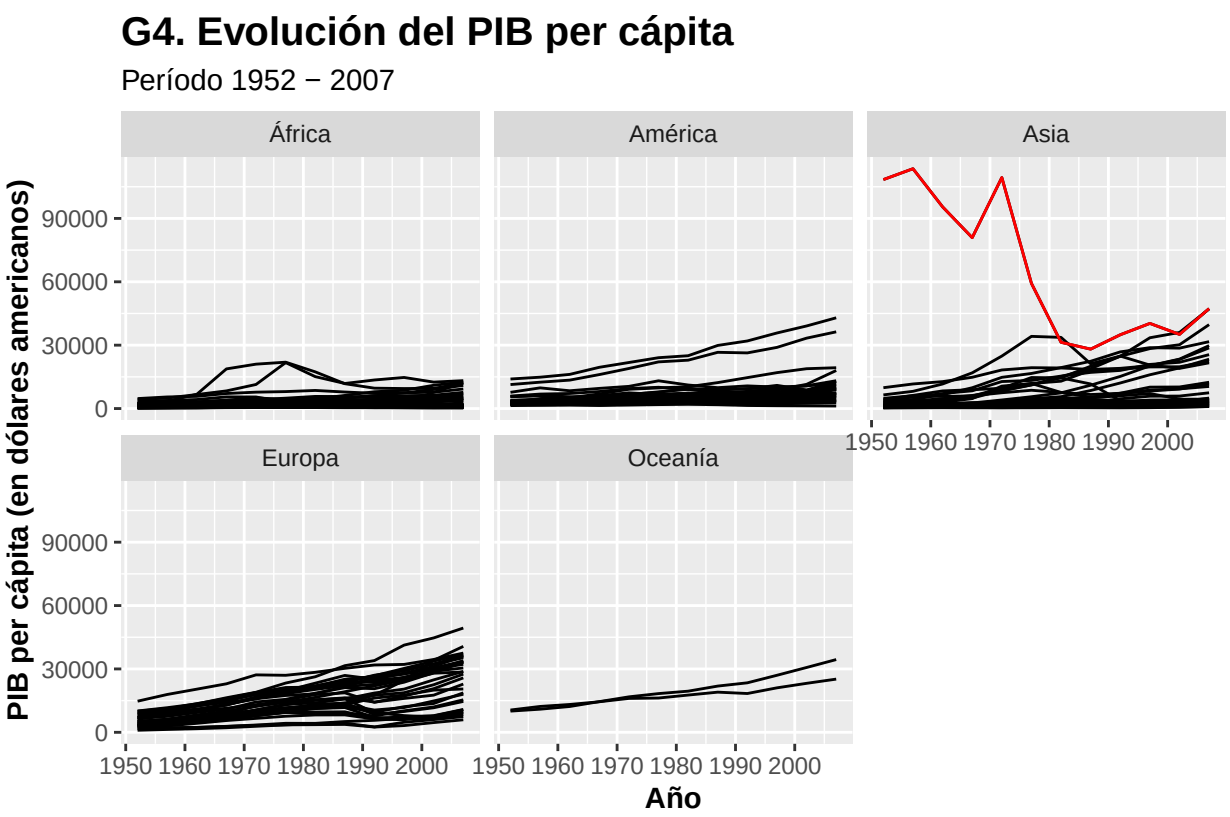
Relación entre la esperanza de vida al nacer y el PIB per cápita, distinguiendo según continente.

En el gráfico G3 podemos observar una relación positiva entre la esperanza de vida al nacer y el PIB per cápita en todos los continentes. Esto significa que a mayor PIB per cápita se puede esperar que la esperanza de vida promedio al nacer sea mayor.

Podríamos decir que un modelo de regresión lineal proporciona un ajuste muy bueno en Oceanía, pero quizás no tanto en el resto de los continentes, donde la relación entre las variables es más difusa.

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.***

```
gapminder %>%
    ggplot(aes(x = year, y = gdpPercap, group = country)) +
    geom_line() + facet_wrap(vars(continent), labeller = as_labeller(c(Africa = "África",
    Americas = "América", Asia = "Asia", Europe = "Europa",
    Oceania = "Oceanía"))) + labs(x = "Año", y = "PIB per cápita (en dólares americanos)",
    title = "G4. Evolución del PIB per cápita", subtitle = "Período 1952 - 2007",
    caption = "Relación entre el PIB per cápita y el tiempo, distinguiendo según continente y país.") +
    theme(plot.title = element_text(face = "bold",
        size = 15), axis.title = element_text(face = "bold")) +
    geom_line(data = filter(gapminder, country == "Kuwait"),
        color = "red")
```



Relación entre el PIB per cápita y el tiempo, distinguiendo según continente y país.

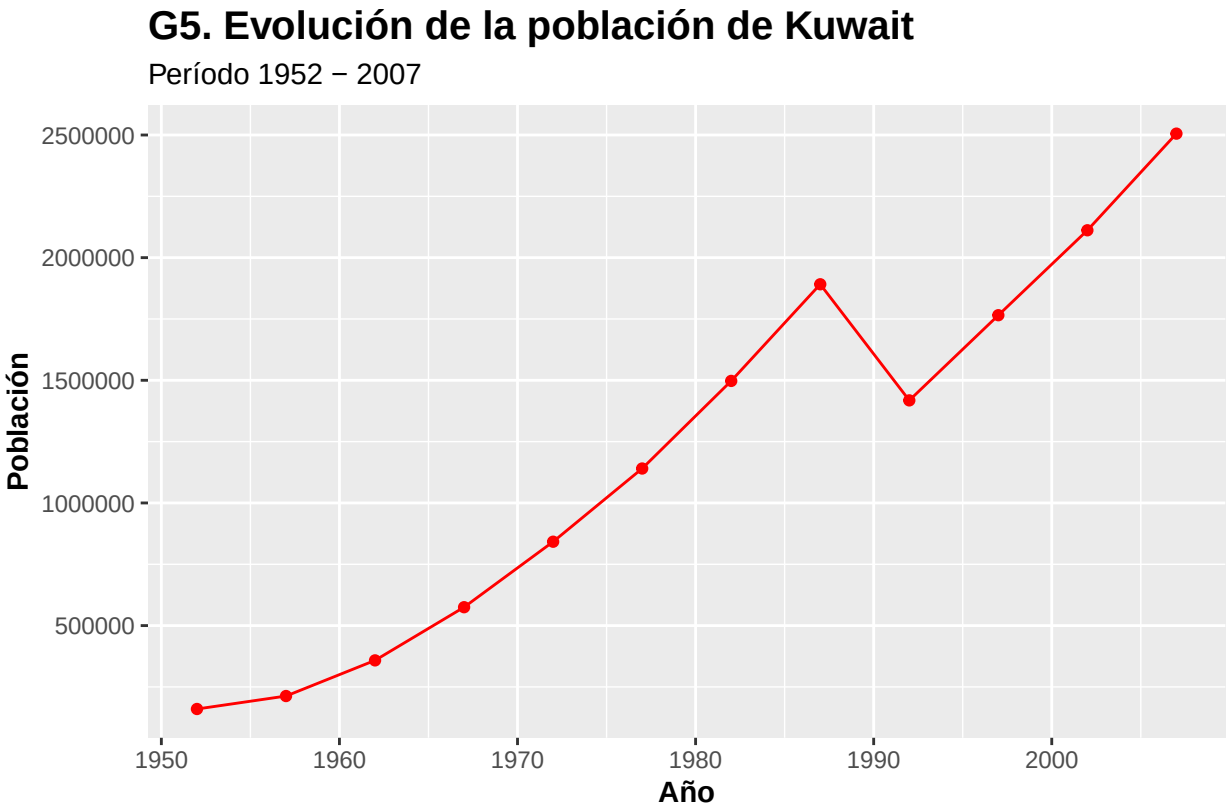
Lo primero que llama nuestra atención es el país de Asia que presenta una tendencia totalmente opuesta al resto de los países y que, adicionalmente, ostenta valores significativamente altos de PIB per cápita. Hay evidencia que nos hace sospechar que se trata de un *outlier* o dato atípico, por lo que sería conveniente investigar de qué país se trata para intentar explicar su comportamiento.

```
gapminder %>%
    filter(gdpPercap == max(gdpPercap))
```

```
## # A tibble: 1 x 6
##   country continent  year lifeExp    pop gdpPercap
##   <fct>   <fct>     <int>   <dbl>  <int>     <dbl>
## 1 Kuwait  Asia       1957    58.0 212846   113523.
```

El resultado anterior nos permite identificar que Kuwait es el *outlier*.

```
gapminder %>%
    filter(country == "Kuwait") %>%
    ggplot(aes(x = year, y = pop)) + geom_point(color = "red") +
    geom_line(color = "red") + labs(x = "Año", y = "Población",
    title = "G5. Evolución de la población de Kuwait",
    subtitle = "Período 1952 - 2007", caption = "Relación entre la población y el tiempo.") +
    theme(plot.title = element_text(face = "bold",
        size = 15), axis.title = element_text(face = "bold"))
```



Relación entre la población y el tiempo.

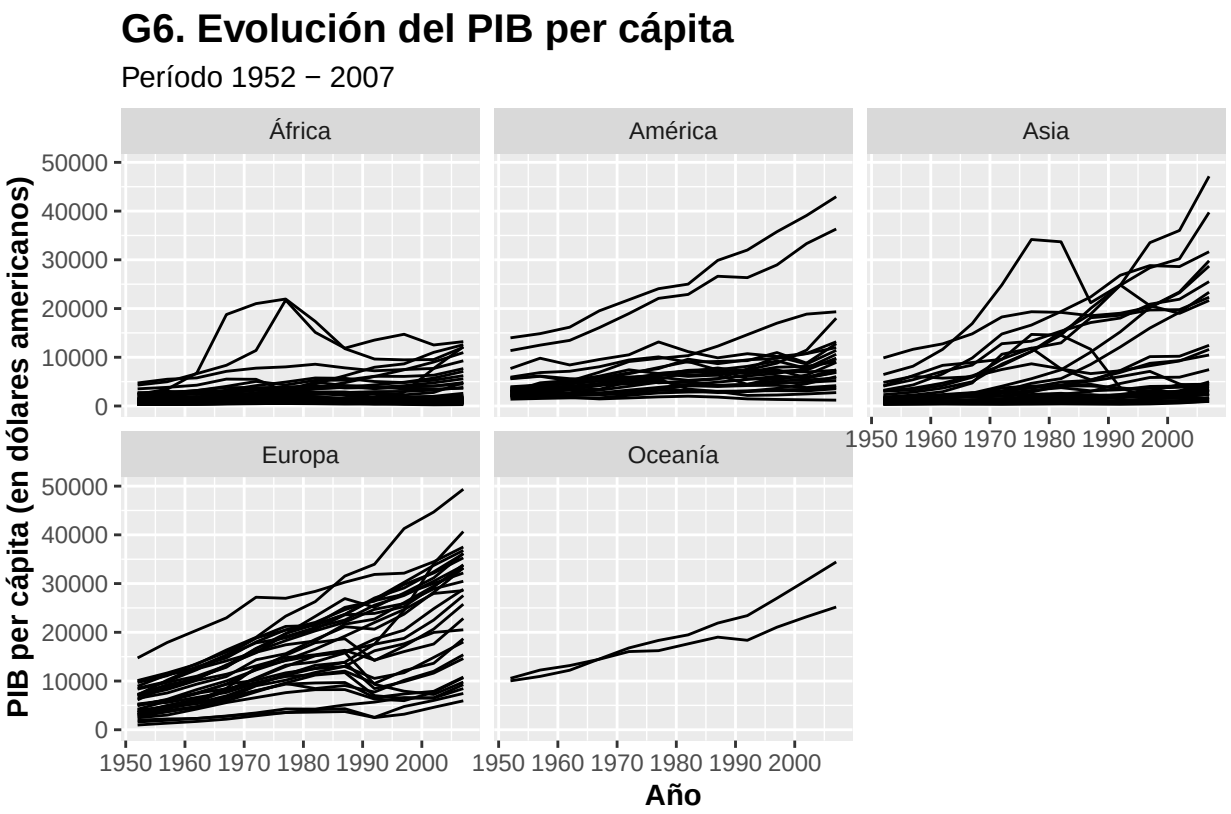
```
gapminder %>%
    filter(country == "Kuwait") %>%
    summarise(year, pop, gdpPercap) %>%
    arrange(year)
```

```
## # A tibble: 12 x 3
##     year     pop gdpPercap
##    <int>   <int>     <dbl>
##  1  1952  160000   108382.
##  2  1957  212846   113523.
##  3  1962  358266    95458.
##  4  1967  575003    80895.
##  5  1972  841934   109348.
##  6  1977 1140357    59265.
##  7  1982 1497494    31354.
##  8  1987 1891487    28118.
##  9  1992 1418095    34933.
## 10  1997 1765345    40301.
## 11  2002 2111561    35110.
## 12  2007 2505559    47307.
```

Al analizar los datos encontramos un comportamiento interesante en la población del país. Vemos que la cantidad de habitantes pasó de tan solo 160.000 en 1952 a 2.505.559 en 2007, representando un incremento sustancial del 1.465,97%. Esta podría ser una explicación al comportamiento que observamos en su PIB per cápita.

Teniendo en cuenta que Kuwait es un *outlier*, podría interesarnos ver la evolución del PIB per cápita, pero esta vez excluyendo a Kuwait.

```
gapminder %>%
    filter(country != "Kuwait") %>%
    ggplot(aes(x = year, y = gdpPercap, group = country)) +
    geom_line() + facet_wrap(vars(continent), labeller = as_labeller(c(Africa = "África",
    Americas = "América", Asia = "Asia", Europe = "Europa",
    Oceania = "Oceanía"))) + labs(x = "Año", y = "PIB per cápita (en dólares americanos)",
    title = "G6. Evolución del PIB per cápita", subtitle = "Período 1952 - 2007",
    caption = "Relación entre el PIB per cápita y el tiempo, distinguiendo según continente y país.") +
    theme(plot.title = element_text(face = "bold",
        size = 15), axis.title = element_text(face = "bold"))
```



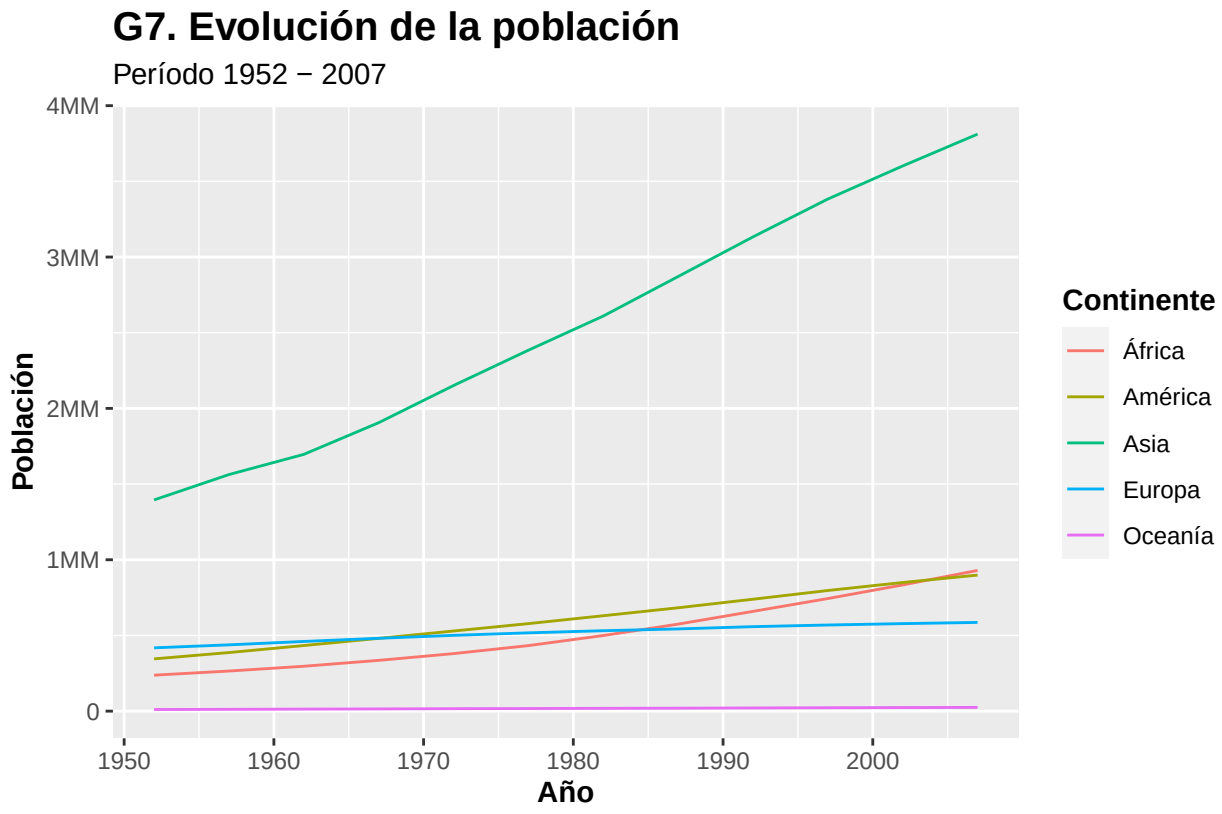
Este gráfico permite apreciar en mayor detalle las diferencias entre los países. Sin embargo, observamos que la gran mayoría siguen una tendencia creciente a lo largo del tiempo, con diferencias en la tasa incremental o presencia de picos en determinados períodos.

Cabe destacar el caso de África, donde encontramos países más estancados y algunos países que experimentaron períodos de auge seguidos de una reversión a la normalidad.

También encontramos países estancados en América y Asia, contrastando nuevamente con Europa y Oceanía, donde se puede apreciar mayor estabilidad característica de los países primermundistas.

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.***

```r
gapminder %>%
    group_by(continent, year) %>%
    summarise(sum_pop = sum(pop)) %>%
    ggplot(aes(x = year, y = sum_pop, color = continent)) +
    geom_line() + scale_color_brewer(palette = "Dark2") +
    labs(x = "Año", y = "Población", color = "Continente",
        title = "G7. Evolución de la población", subtitle = "Período 1952 - 2007",
        caption = "Relación entre la población y el tiempo, distinguiendo según continente.") +
    scale_color_discrete(labels = c(Africa = "África",
        Americas = "América", Europe = "Europa", Oceania = "Oceanía")) +
    theme(plot.title = element_text(face = "bold",
        size = 15), axis.title = element_text(face = "bold"),
        legend.title = element_text(face = "bold")) +
    scale_y_continuous(breaks = c(0, 1e+09, 2e+09,
        3e+09, 4e+09), labels = c("0", "1MM", "2MM",
        "3MM", "4MM"))
```



Relación entre la población y el tiempo, distinguiendo según continente.

Intuitivamente, Asia es el continente que presenta mayor población y, a su vez, mayor crecimiento poblacional a lo largo del tiempo.

También podemos observar que la población de Oceanía permanece prácticamente constante y la población de Europa presenta un incremento moderado respecto a 1952.

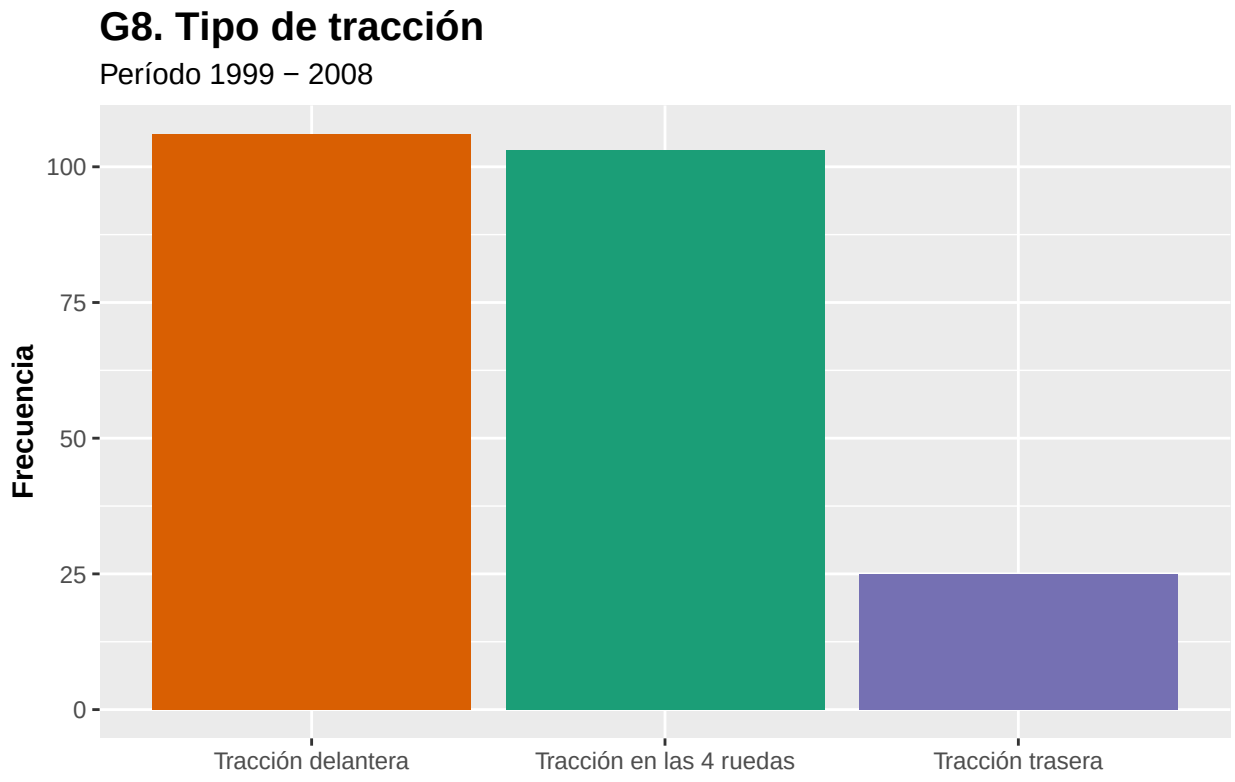
Por otro lado, América y África triplicaron su población en el período, alcanzando en 2007 una población similar entre sí.

Ejercicio 2

1. *Con los datos `mpg` que se encuentran disponible en `ggplot2` hacer un gráfico de barras para la variable `drv` con las siguientes características:*

- *Las barras tienen que estar coloreadas por `drv`.*
- *Incluir usando `labs()` el nombre de los ejes y título informativo.*
- *Usá la paleta de colores `Dark2`, mirá la ayuda de `scale_colour_brewer()`.*

```
mpg %>%
    ggplot() + geom_bar(aes(x = fct_infreq(drv), fill = drv)) +
    scale_fill_brewer(palette = "Dark2") + labs(x = "",
    y = "Frecuencia", title = "G8. Tipo de tracción",
    subtitle = "Período 1999 - 2008", caption = "Comparación del tipo de tracción para 38 modelos de au
    theme(plot.title = element_text(face = "bold",
        size = 15), axis.title = element_text(face = "bold"),
        legend.position = "none") + scale_x_discrete(labels = c('4' = "Tracción en las 4 ruedas",
    f = "Tracción delantera", r = "Tracción trasera"))
```



Comparación del tipo de tracción para 38 modelos de auto.

Del gráfico G8 se desprende que el tipo de tracción más frecuente es la delantera, seguida de cerca por la tracción en las cuatro ruedas y, por último y bastante menos frecuente, la tracción trasera.

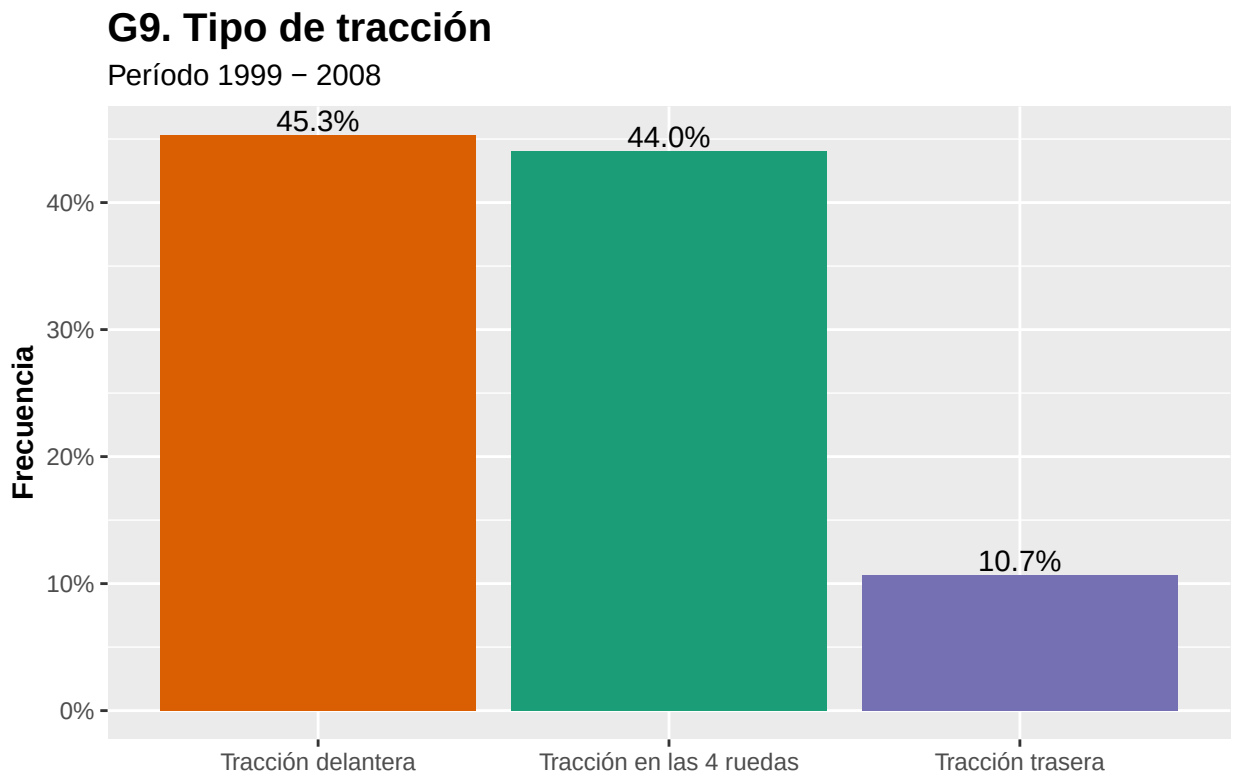
2. ***Usando como base el gráfico anterior:***

- ***Incluir en el eje y porcentaje en vez de conteos.***

- *Usando `scale_y_continuous()` cambiar la escala del eje y a porcentajes.*
- *Usando `geom_text()` incluir texto con porcentajes arriba de cada barra.*

```
mpg %>%
    group_by(drv) %>%
    summarise(count = n()) %>%
    mutate(drv = factor(drv), percentage = count/sum(count)) %>%

ggplot(aes(x = reorder(drv, -percentage), y = percentage,
    fill = drv)) + geom_bar(stat = "identity") + geom_text(aes(label = percent(percentage)),
    vjust = -0.2) + scale_fill_brewer(palette = "Dark2") +
    labs(x = "", y = "Frecuencia", title = "G9. Tipo de tracción",
        subtitle = "Período 1999 - 2008", caption = "Comparación del tipo de tracción para 38 modelos d
    theme(plot.title = element_text(face = "bold",
        size = 15), axis.title = element_text(face = "bold"),
        legend.position = "none") + scale_x_discrete(labels = c('4' = "Tracción en las 4 ruedas",
    f = "Tracción delantera", r = "Tracción trasera")) +
    scale_y_continuous(label = scales::percent_format(accuracy = 1))
```



Comparación del tipo de tracción para 38 modelos de auto.

Ejercicio 3

Los datos que vamos a utilizar en este ejercicio están disponibles en el catálogo de datos abiertos Uruguay https://catalogodatos.gub.uy. Los datos que seleccioné son sobre las emisiones de dióxido de carbono (CO2) correspondientes a las actividades de quema de los combustibles

en las industrias de la energía y los sectores de consumo. Se incluyen también emisiones de CO2 provenientes de la quema de biomasa y de bunkers internacionales, las cuales se presentan como partidas informativas ya que no se consideran en los totales. En el siguiente link se encuentran los datos y los meta datos con información que describe la base de datos https://catalogodatos.gub.uy/dataset/miem-emisiones-de-co2-por-sector.

Por simplicidad te damos los datos restructurados (veremos como se hace más adelante en el curso), el archivo se llama datos_emisión.csv, contiene tres columnas AÑO, fuente y emisión.

1. *Leer los datos usando el paquete* `readr` *y la función* `read_csv`*, guardarlos en un objeto llamado* `datos`*.*

```
library(readr)
datos <- read_csv("dato_emision.csv")
```

2. *Usando las funciones de la librería* `dplyr` *obtenga qué fuentes tienen la emisión máxima. Recuerde que TOTAL debería ser excluido para esta respuesta así como los subtotales.*

```
datos %>%
    filter(fuente != "TOTAL" & fuente != "S_C" & fuente !=
        "I_E") %>%
    group_by(fuente) %>%
    summarise(max_emision = max(emision)) %>%
    arrange(desc(max_emision))
```

```
## # A tibble: 10 x 2
##    fuente max_emision
##    <chr>        <dbl>
##  1 Q_B          9070.
##  2 T            3734
##  3 CE_SP        2925.
##  4 BI           1803.
##  5 I             894.
##  6 A_P_M         602.
##  7 R             500
##  8 CP            482.
##  9 C_S_SP        163.
## 10 NI             NA
```

La quema de biomasa representa la mayor fuente de emisión de CO2.

3. *¿En qué año se dio la emisión máxima para la fuente que respondió en la pregunta anterior?*

```
datos %>%
    filter(fuente == "Q_B") %>%
    arrange(desc(emision)) %>%
    head(1)
```

```
## # A tibble: 1 x 3
##     AÑO fuente emision
##   <dbl> <chr>    <dbl>
## 1  2017 Q_B      9070.
```

En el año 2017.

4. ***Usando las funciones de la librería `dplyr` obtenga las 5 fuentes, sin incluir TOTAL ni subtotales, qué tienen un valor medio de emisión a lo largo de todos los años más grandes.***

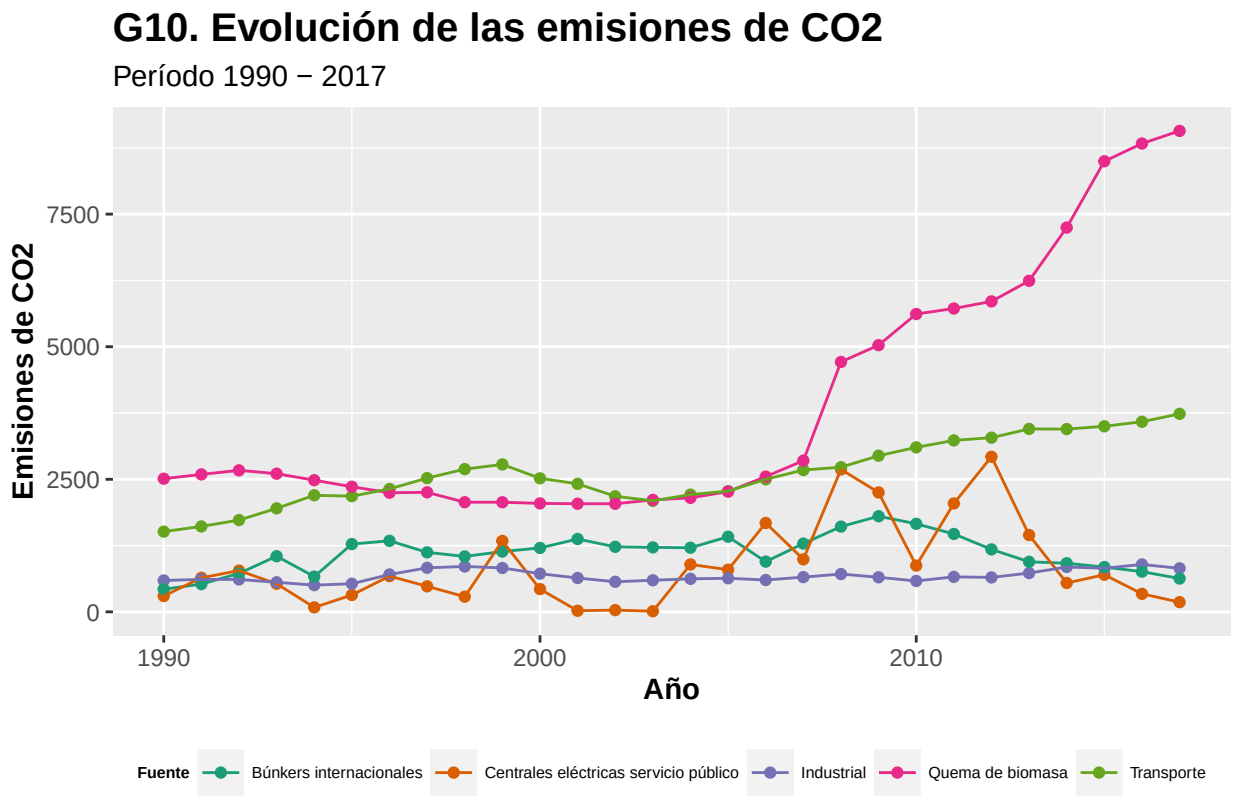
```
datos %>%
    filter(fuente != "TOTAL" & fuente != "S_C" & fuente !=
        "I_E") %>%
    group_by(fuente) %>%
    summarise(mean_emision = mean(emision)) %>%
    arrange(desc(mean_emision)) %>%
    head(5)
```

```
## # A tibble: 5 x 2
##   fuente mean_emision
##   <chr>         <dbl>
## 1 Q_B           3883.
## 2 T             2621.
## 3 BI            1107.
## 4 CE_SP          867.
## 5 I              680.
```

Las fuentes con mayor emisión media son la quema de biomasa, el transporte, los búnkers internacionales, las centrales eléctricas servicio público y el sector industrial, en ese orden.

5. ***Usando ggplot2 realice un gráfico de las emisiones a lo largo de los años para cada fuente. Utilice dos elementos geométricos, puntos y lineas. Seleccione para dibujar solamente las 5 fuentes que a lo largo de los años tienen una emisión media mayor que el resto (respuesta de la pregunta 5). 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.***

```
datos %>%
    filter(fuente == "Q_B" | fuente == "T" | fuente ==
        "BI" | fuente == "CE_SP" | fuente == "I") %>%
    ggplot(aes(x = AÑO, y = emision, color = fuente)) +
    geom_point() + geom_line() + labs(x = "Año", y = "Emisiones de CO2",
    color = "Fuente", title = "G10. Evolución de las emisiones de CO2",
    subtitle = "Período 1990 - 2017", caption = "Relación entre las emisiones de CO2 y el tiempo, distin
    theme(plot.title = element_text(face = "bold",
        size = 15), axis.title = element_text(face = "bold"),
        legend.title = element_text(face = "bold",
            size = 6), legend.position = "bottom",
        legend.text = element_text(size = 6)) + scale_color_brewer(palette = "Dark2",
    labels = c(BI = "Búnkers internacionales", CE_SP = "Centrales eléctricas servicio público",
        I = "Industrial", Q_B = "Quema de biomasa",
        T = "Transporte"))
```



Relación entre las emisiones de CO2 y el tiempo, distinguiendo según fuente de emisión.

Podemos ver que los búnkers internacionales y las industrias presentan una trayectoria más o menos estable a lo largo del tiempo, mientras que las centrales eléctricas denotan mayor volatilidad, con picos a lo largo de todo el período. Por su parte, el transporte y la quema de biomasa no solo son las mayores fuentes de emisión de CO2, sino que además presentan una trayectoria creciente a lo largo del tiempo.

En particular, resulta un tanto alarmante el incremento de emisiones de CO2 por quema de biomasa en los últimos años. No solamente por la emisión en sí misma, sino por el incremento en la cantidad de biomasa quemada que esto implica.

Los principales materiales que se utilizan para la quema de biomasa son madera y basura. Por lo tanto, un incremento en la cantidad de biomasa quemada es debida a un incremento en la madera talada o bien a un incremento de la basura generada. Ambos representan situaciones no deseadas, pero el incremento de la madera talada es de especial importancia ya que son los bosques los encargados de absorber el CO2 presente en la atmósfera. Al talar madera para la quema de biomasa no solo estamos incrementando las emisiones de CO2, sino que estamos disminuyendo la absorción del mismo.

Adicionalmente, el CO2 no es la única sustancia que se emite a la atmósfera debido a la quema de biomasa, sino que también se emiten sustancias como:

- **Sustancias emitidas por la quema de madera:** sustancias catalogadas de cancerígenas como el benzopireno y ciertos hidrocarburos.
- **Sustancias emitidas por la quema de basura:** sustancias altamente contaminantes debido a la presencia de plásticos y cloro como las dioxinas, furanos y el ácido clorhídrico (sustancias tóxicas).
- **Sustancias emitidas por el proceso de combustión:** sustancias tóxicas y/o cancerígenas como el óxido de nitrógeno, monóxido de carbono, óxido de azufre o ácido sulfúrico, arsénico, ácido acético, fenol, benceno, tolueno, benzopireno y residuos sólidos y líquidos como las cenizas.

Como si esto fuera poco, durante el proceso se utilizan sustancias que pueden contaminar agua y suelo, debido a las sustancias tóxicas entre los compuestos que constituyen las cenizas (plomo y cadmio).

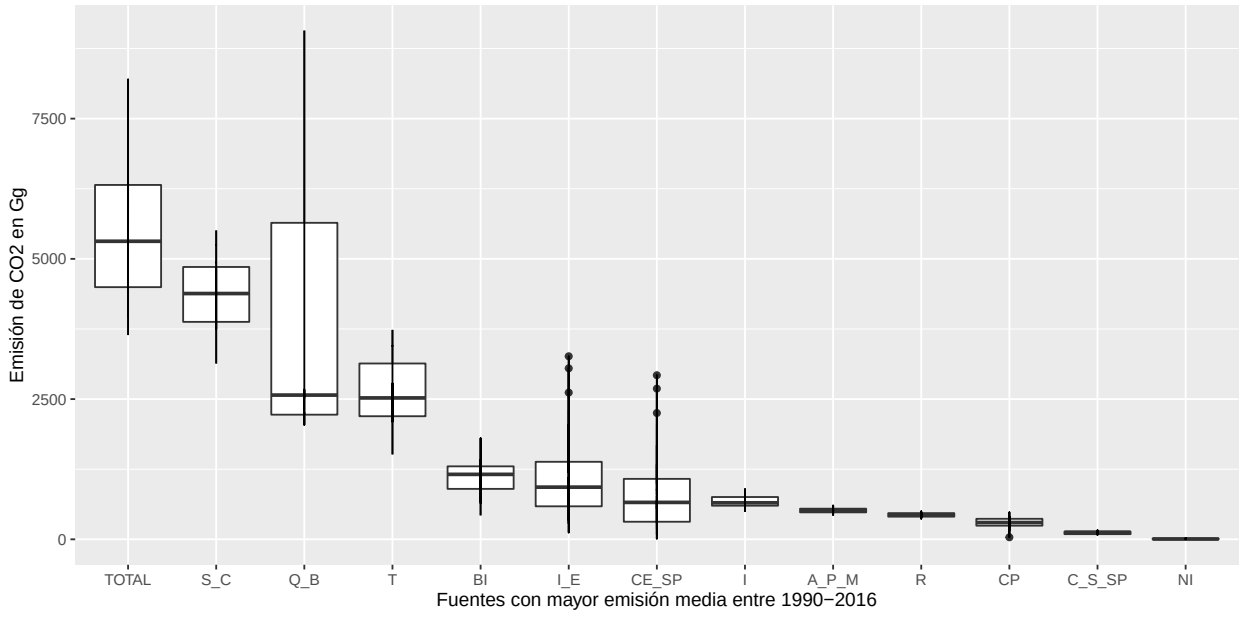
Esta batería de sustancias no solo representan un riesgo para el medioambiente, sino también para la salud de las personas, ya que las sustancias emitidas pasan al medio (agua, suelo, aire) y a los alimentos, llegando fácilmente a las personas.

Dentro de las principales afectaciones a la salud derivadas de absorber estos contaminantes destacan:

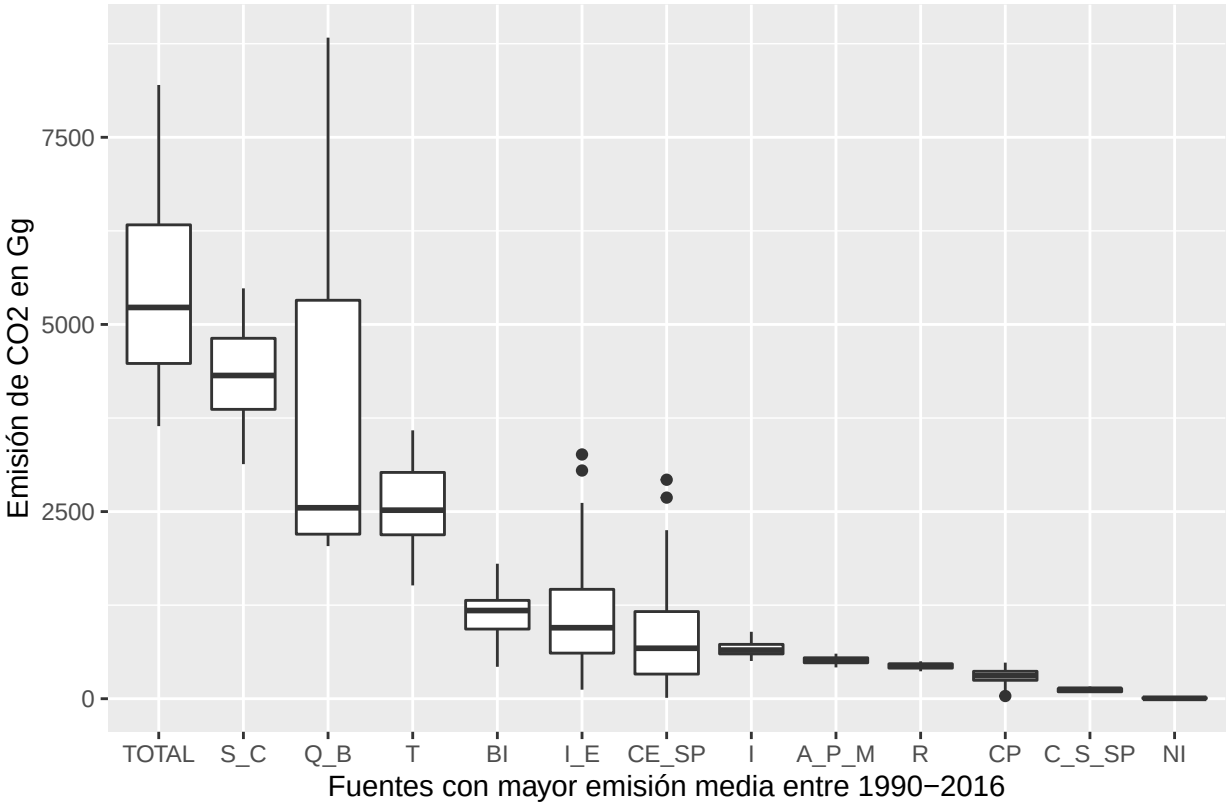
- Afecciones respiratorias como bronquitis o asma.
- Enfermedades cardiovasculares como infartos de miocardio y accidentes cerebrovasculares (ACVs).
- Efectos neurológicos como párkinson o alzhéimer.
- Distintos tipos de cáncer como cáncer de pulmón, de mama o leucemia.
- Efectos endócrinos como diabetes.
- Mortalidad prematura.

Con base en estos resultados podemos afirmar que esta fuente de emisión debería ser el principal foco de actuación de toda política que busque reducir la contaminación ambiental por emisiones de CO2 y proteger la salud de la población.

6. ***Replique el siguiente gráfico usando ggplot2. Incluir un `caption` en la figura con algún comentario de interés que describa el gráfico.***




```
datos %>%
    filter(AÑO < 2017) %>%
    ggplot(aes(x = reorder(fuente, -emision, median),
        y = emision)) + geom_boxplot() + labs(x = "Fuentes con mayor emisión media entre 1990-2016",
    y = "Emisión de CO2 en Gg", caption = "Distribución de las emisiones de CO2 según las distintas fue
```

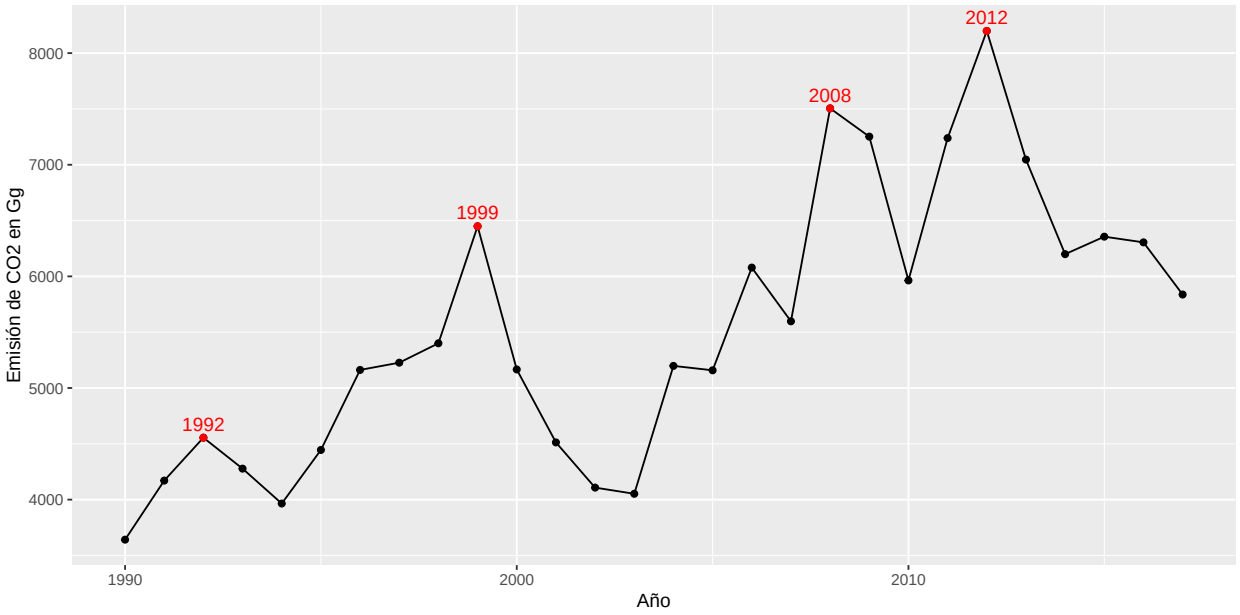


Distribución de las emisiones de CO2 según las distintas fuentes de emisión.

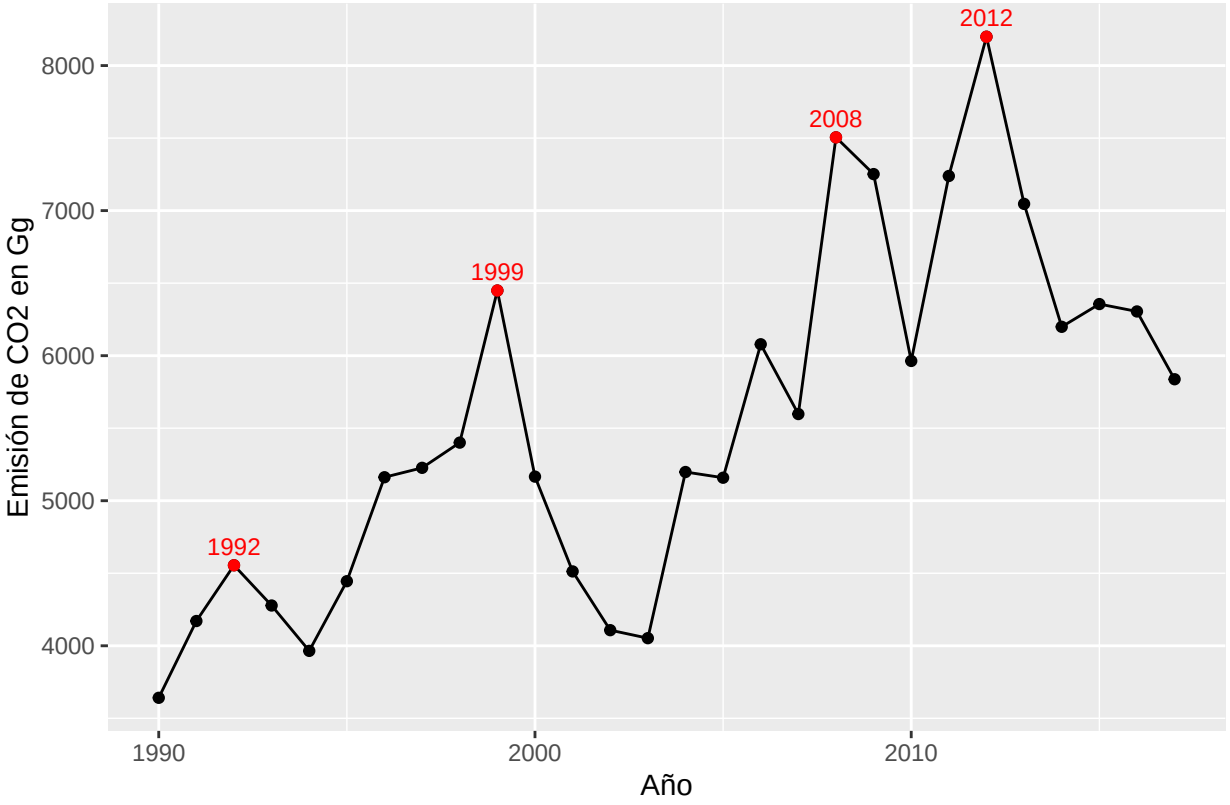
En el gráfico anterior podemos apreciar que nuevamente destaca la emisión por quema de biomasa, ya que la caja está situada por encima de las demás fuentes de emisión (a excepción del total y el subtotal del sector de consumo).

Vemos que presenta una caja mucho más grande que las demás, lo cual implica que el 50% de los datos están más dispersos. A su vez, la mayoría de las observaciones están por encima de la mediana, lo cual se corresponde con el incremento de los últimos años que veíamos en el gráfico G10. Esto también explica que el bigote superior sea tan largo en comparación con las otras fuentes de emisión.

7. ***Usando la librería `ggplot2` y `ggpmisc` replique el siguiente gráfico de las emisiones totales entre 1990 y 2016. Los puntos rojos indican los máximos locales o picos de emisión de CO2 en Gg. Use `library(help = ggpmisc)` para ver todas las funciones de la librería `ggpmisc` e identificar cual o cuales necesita para replicar el gráfico. Incluir un `caption` en la figura con algún comentario de interés que describa el gráfico.***


2012
2008
1999
1992
8000
7000
6000
5000
4000
Emisión de CO2 en Gg
1990
2000
2010
Año

```r
datos %>%
    filter(fuente == "TOTAL") %>%
    ggplot(aes(x = AÑO, y = emision)) + geom_point() +
    geom_line() + labs(x = "Año", y = "Emisión de CO2 en Gg",
    caption = "Relación entre el total de emisiones de CO2 y el tiempo.") +
    stat_peaks(color = "red", geom = "point") + stat_peaks(color = "red",
    geom = "text", , vjust = -0.6, size = 3)
```



Relación entre el total de emisiones de CO2 y el tiempo.

En el gráfico anterior vemos que las emisiones alcanzaron un pico en 1999 seguido de una caída abrupta, probablemente relacionada con la firma del Protocolo de Kioto en ese mismo año, con el objetivo de reducir las emisiones de gases de efecto invernadero a nivel internacional.

Sin embargo, observamos que a partir de 2003 las emisiones comenzaron a aumentar hasta alcanzar un récord histórico a nivel mundial en 2008. A partir de este año también encontramos una caída pronunciada debida a varios factores de los cuales se destacan:

- El alto precio del CO2 y del petróleo que redujeron el uso del transporte.
- La crisis económica que provocó un descenso en la producción industrial.
- El desplazamiento hacia fuentes alternativas de energía como la eólica y la hidráulica.

Las emisiones de CO2 alcanzaron un nuevo récord en 2012, año en el cual finalizaba el primer período de compromiso del Protocolo de Kioto. Adicionalmente, la NASA catalogó a este año como el noveno más caluroso desde 1880, lo cual alertó a la población sobre los efectos del cambio climático. En 2013 comenzó el segundo período de compromiso del Protocolo, lo cual sumado a la mayor conscientización ambiental repercutió en la disminución de las emisiones de CO2.

Ejercicio 4

Los datos que vamos a utilizar en este ejercicio son una muestra de datos a nivel nacional sobre abandono escolar en los años 2016.

Table 1: Variables en **muestra.csv**

Variable	Descripción
documento	Cédula de Identidad del alumno
nro_doc_centro_educ	Liceo que concurre el alumno en 2016
nombre_departamento	Nombre del Departamento del centro educativo
grupo_desc	Grupo del alumno en 2016
coberturaT	Cobertura en el primer semestre de 2016
Centro_Grupo	Liceo y grupo del alumno en 2016
cl	Cluster - contexto sociocultural del liceo en 1016
Grado_2016_UE	Grado del alumno en el 2016 según UE
Grado2013	Grado del alumno en 2013 según CRM
Grado2014	Grado del alumno en 2014 según CRM
Grado2015	Grado del alumno en 2015 según CRM
Grado 2016	Grado del alumno en 2016 según CRM
Sexo	Sexo del alumno
Fecha.nacimiento	Fecha de nacimiento del alumno
Grupo_UE_2017	Grupo del alumno en 2017
inasistencias	cantidad de inasistencias en el primer semestre de 2016
asistencias	cantidad de asistencias en el primer semestre de 2016

En el Cuadro 1 se presentan las variables en el conjunto de datos muestra.csv.

Este ejercicio tiene como objetivo que realice tres preguntas de interés que le surgen como parte del análisis exploratorio de datos utilizando todo lo aprendido en el curso.

Debe plantear 3 preguntas orientadoras y visualizaciones apropiadas para responderlas. La exploración deberá contener las preguntas a responder sus respuestas con el correspondiente resumen de información o visualización. Incluya en su exploración el análisis de la variabilidad tanto de variables cuantitativas como cualitativas y covariaciones entre las mismas. Recuerde que en las visualizaciones, 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 y lo que ve en el mismo.

```
muestra <- read_csv("muestra.csv")
```

Preguntas orientadoras:

1. ¿Cuál es el departamento con la mayor tasa de inasistencias en Uruguay?
2. ¿Existe una relación entre la tasa de inasistencias y el nivel de abandono?
3. ¿Existen diferencias en el nivel de abandono según contexto sociocultural del liceo? ¿Y según sexo?

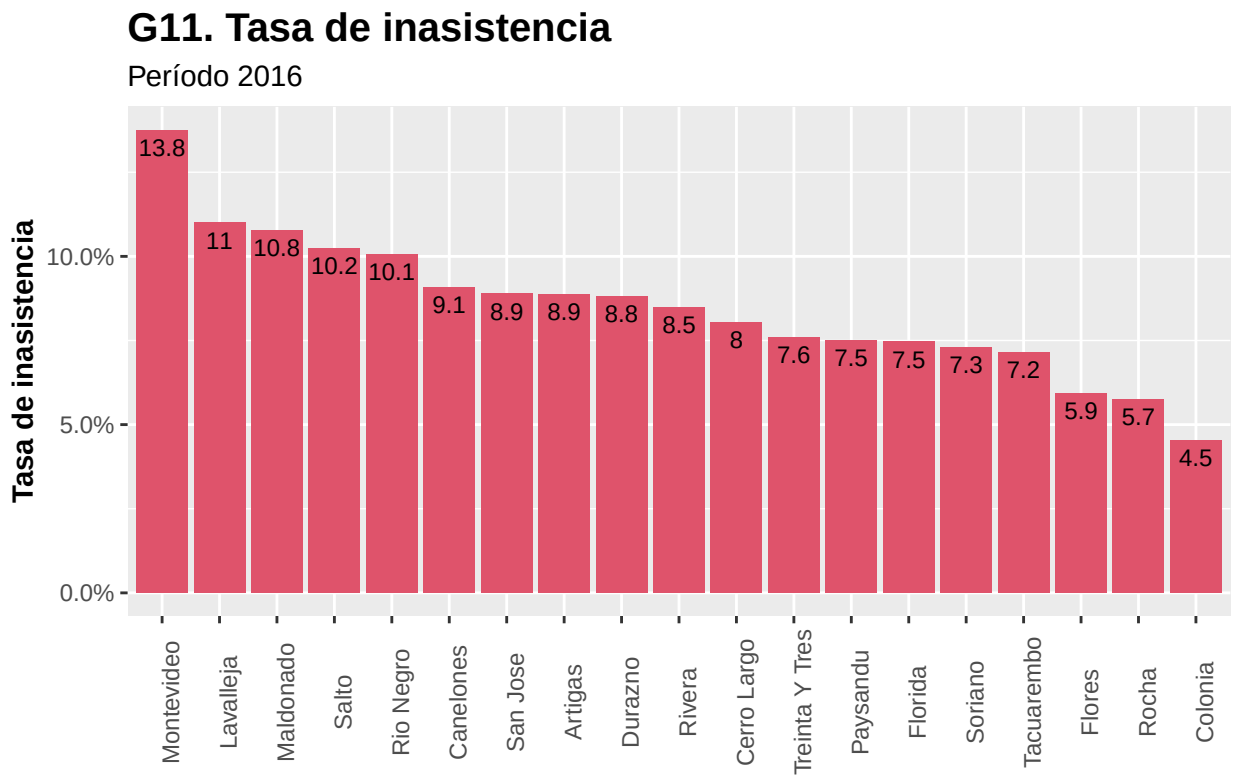
Pregunta 1:

¿Cuál es el departamento con la mayor tasa de inasistencias en Uruguay?

```r
preg1 <- muestra %>%
    group_by(nombre_departamento) %>%
    summarise(total_inasistencias = sum(inasistencias),
        total_asistencias = sum(asistencias), tasa_inasistencia = total_inasistencias/(total_asistencias
            total_inasistencias), tasa_asistencia = 1 -
            tasa_inasistencia)
preg1
```

```
## # A tibble: 19 x 5
##    nombre_departamento total_inasistencias total_asistencias tasa_inasistencia
##    <chr>                             <dbl>             <dbl>              <dbl>
##  1 Artigas                             449              4615             0.0887
##  2 Canelones                          2914             29169             0.0908
##  3 Cerro Largo                         546              6251             0.0803
##  4 Colonia                             349              7339             0.0454
##  5 Durazno                             274              2837             0.0881
##  6 Flores                              110              1745             0.0593
##  7 Florida                             353              4379             0.0746
##  8 Lavalleja                           514              4160             0.110
##  9 Maldonado                           966              7991             0.108
## 10 Montevideo                         8304             52067             0.138
## 11 Paysandu                            545              6725             0.0750
## 12 Rio Negro                           379              3384             0.101
## 13 Rivera                              946             10195             0.0849
## 14 Rocha                               312              5120             0.0574
## 15 Salto                               911              7979             0.102
## 16 San Jose                            716              7339             0.0889
## 17 Soriano                             381              4842             0.0729
## 18 Tacuarembo                          540              7005             0.0716
## 19 Treinta Y Tres                      263              3196             0.0760
## # ... with 1 more variable: tasa_asistencia <dbl>
```

```r
preg1 %>%
    ggplot(aes(x = reorder(nombre_departamento, -tasa_inasistencia),
        y = tasa_inasistencia)) + geom_col(fill = 2) +
    labs(x = "", y = "Tasa de inasistencia", title = "G11. Tasa de inasistencia",
        subtitle = "Período 2016", caption = "Tasa de inasistencia según departamento.") +
    theme(plot.title = element_text(face = "bold",
        size = 15), axis.title = element_text(face = "bold"),
        axis.text.x = element_text(angle = 90)) + scale_y_continuous(label = scales::percent) +
    geom_text(aes(label = round(tasa_inasistencia,
        3) * 100), size = 3, vjust = 1.5)
```



Tasa de inasistencia según departamento.

Montevideo es el departamento con una tasa de inasistencia más alta, ascendiendo al 13.8% de los estudiantes de la muestra. Por su parte, Colonia es el departamento que presenta una menor tasa de inasistencia, de solo un 4.5%.

Pregunta 2:

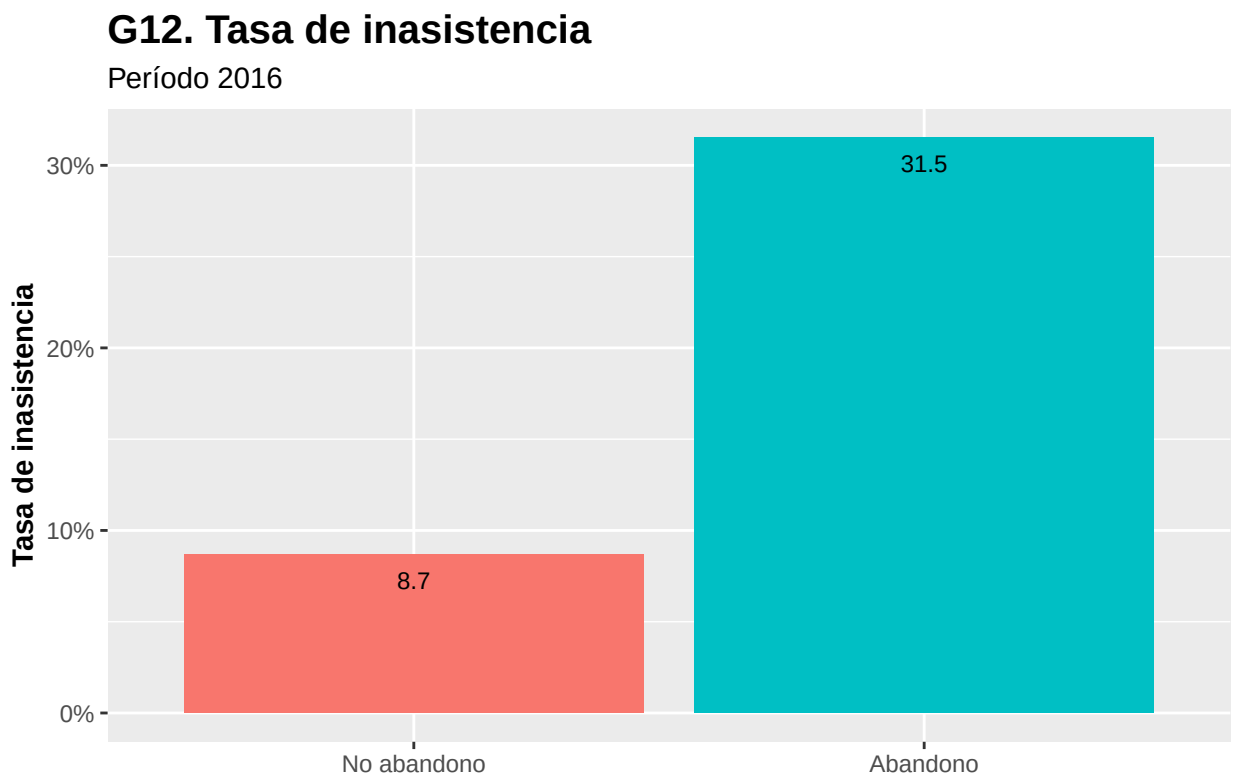
¿Existe una relación entre la tasa de inasistencias y el nivel de abandono?

```
preg2 <- muestra %>%
    group_by(Abandono) %>%
    summarise(total_inasistencias = sum(inasistencias),
        total_asistencias = sum(asistencias), tasa_inasistencia = total_inasistencias/(total_asistencias
            total_inasistencias), tasa_asistencia = 1 -
            tasa_inasistencia)
preg2
```

```
## # A tibble: 2 x 5
##   Abandono total_inasistenci~ total_asistencias tasa_inasistenc~ tasa_asistencia
##      <dbl>              <dbl>              <dbl>            <dbl>           <dbl>
## 1        0              15962             168060           0.0867           0.913
## 2        1               3810               8278           0.315            0.685
```

```
preg2 %>%
    ggplot(aes(x = factor(Abandono), y = tasa_inasistencia,
        fill = factor(Abandono))) + geom_col() + labs(x = "",
    y = "Tasa de inasistencia", title = "G12. Tasa de inasistencia",
    subtitle = "Período 2016", caption = "Relación entre la tasa de inasistencia y el abandono estudian
```

```
    theme(plot.title = element_text(face = "bold",
        size = 15), axis.title = element_text(face = "bold"),
        legend.position = "none") + scale_y_continuous(label = scales::percent) +
    scale_x_discrete(labels = c(`0` = "No abandono",
        `1` = "Abandono")) + geom_text(aes(label = round(tasa_inasistencia,
    3) * 100), size = 3, vjust = 2)
```



Relación entre la tasa de inasistencia y el abandono estudiantil.

En el gráfico G12 encontramos que la tasa de inasistencia es sensiblemente más alta para aquellos estudiantes que abandonan los estudios. Por lo tanto, intuitivamente podríamos decir que existe una relación positiva entre la tasa de inasistencia y el abandono estudiantil.

Pregunta 3:

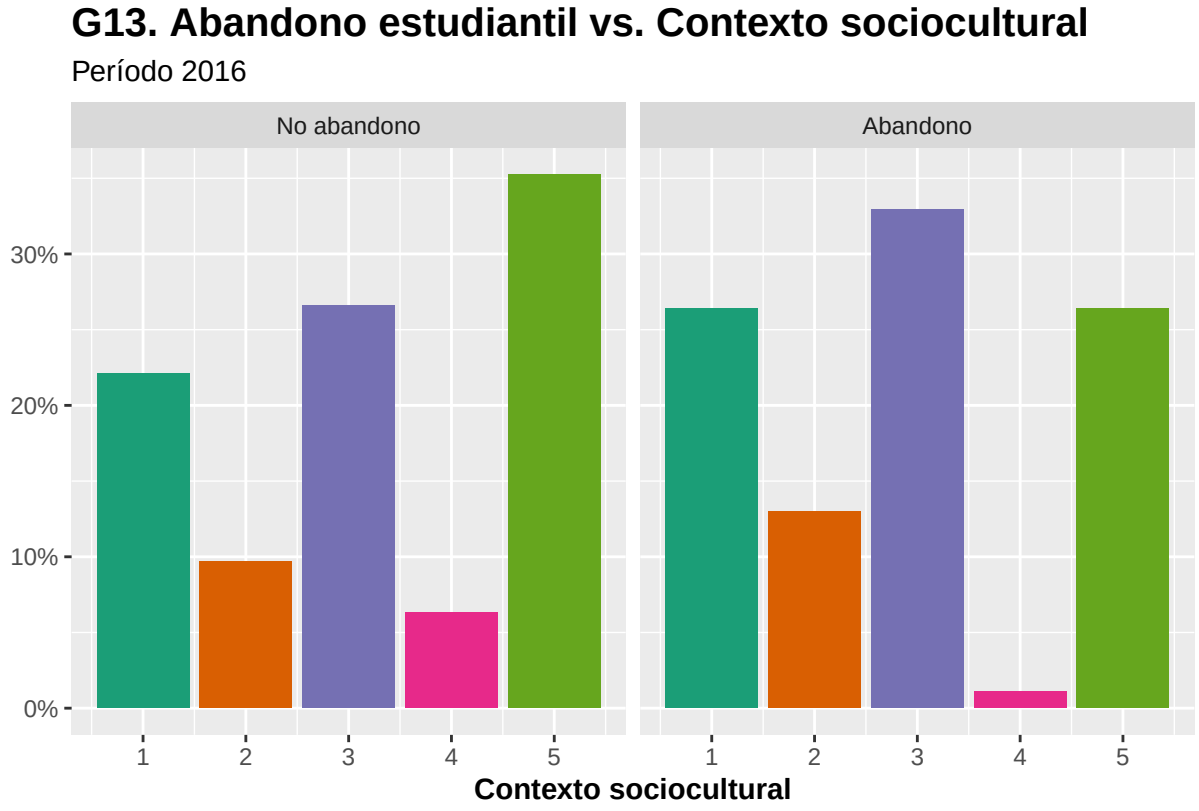
¿Existen diferencias en el nivel de abandono según el contexto sociocultural del liceo? ¿Y según sexo?

```
table(muestra$Abandono, muestra$cl)
```

```
##
##        1    2    3    4    5
##   0  832  366 1000  238 1326
##   1   69   34   86    3   69
```

```
muestra %>%
    group_by(Abandono) %>%
    summarise(cl, count = n()) %>%
```

```
    mutate(Abandono = factor(Abandono), percentage = count/sum(count)) %>%
    group_by(cl) %>%
    ggplot(aes(x = cl, y = percentage, fill = factor(cl))) +
    geom_bar(stat = "identity") + scale_y_continuous(label = scales::percent) +
    facet_wrap(vars(Abandono), labeller = as_labeller(c(`0` = "No abandono",
        `1` = "Abandono"))) + scale_fill_brewer(palette = "Dark2") +
    theme(legend.position = "none") + labs(x = "Contexto sociocultural",
    y = "", title = "G13. Abandono estudiantil vs. Contexto sociocultural",
    subtitle = "Período 2016", caption = "Distribución del abandono y la permanencia estudiantil según
    theme(plot.title = element_text(face = "bold",
        size = 15), axis.title = element_text(face = "bold"),
        legend.position = "none") + scale_y_continuous(label = scales::percent)
```



Distribución del abandono y la permanencia estudiantil según contexto sociocultural.

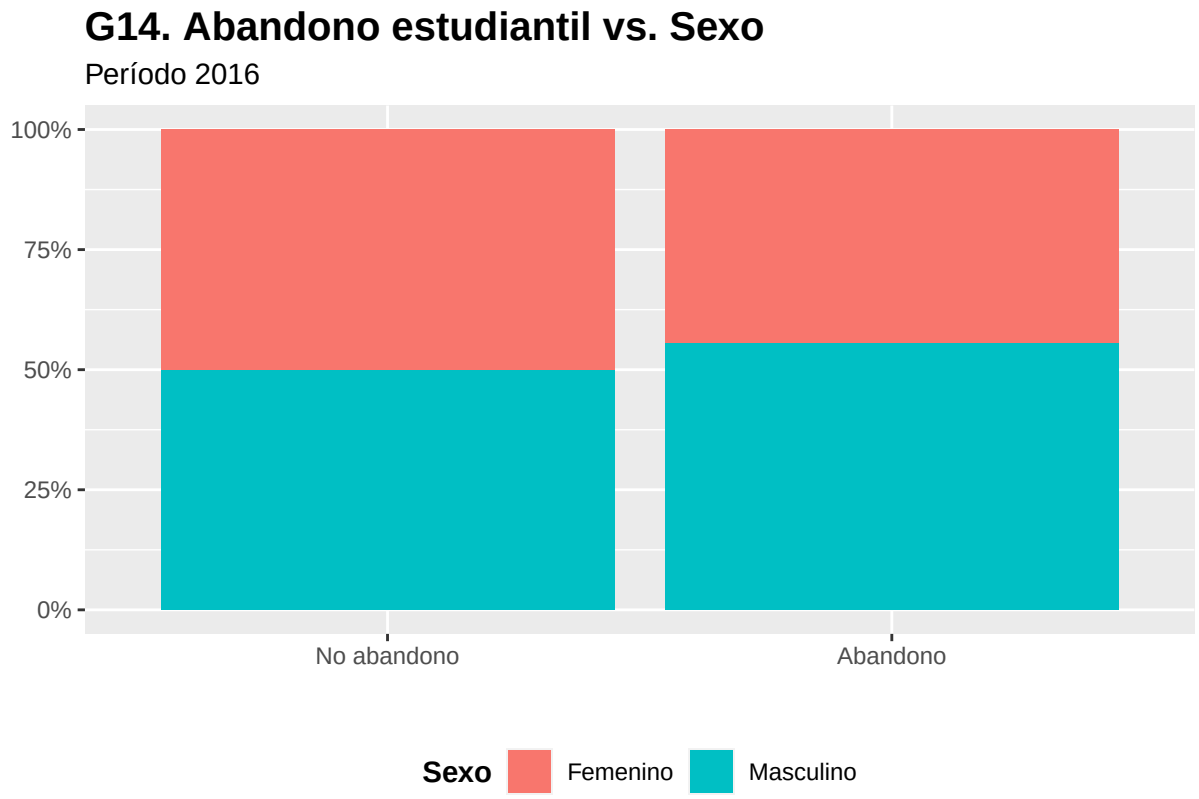
En el gráfico G13 observamos que sí existen diferencias a nivel sociocultural. Se desprende que el contexto 3 es en el cual más estudiantes abandonan, mientras que el contexto 5 es en el cual más estudiantes permanecen en la educación.

Los contextos 1 y 5 son el segundo y tercer contexto con mayor cantidad de abandonos, presentando una cantidad de abandonos similar entre ellos. Por último, el contexto 4 es en el que se registran menos abandonos, pero esto puede deberse a que hay pocas observaciones para este contexto.

```
table(muestra$Sexo, muestra$Abandono)
```

```
##
##        0    1
##   F 1889  116
##   M 1873  145
```

```r
muestra %>%
    ggplot() + geom_bar(aes(x = factor(Abandono), fill = Sexo),
    position = "fill") + labs(x = "", y = "", title = "G14. Abandono estudiantil vs. Sexo",
    subtitle = "Período 2016", caption = "Distribución del abandono y la permanencia estudiantil según s
    theme(plot.title = element_text(face = "bold",
        size = 15), axis.title = element_text(face = "bold"),
        legend.title = element_text(face = "bold"),
        legend.position = "bottom") + scale_y_continuous(labels = scales::percent) +
    scale_fill_discrete(labels = c(F = "Femenino",
        M = "Masculino")) + scale_x_discrete(labels = c(`0` = "No abandono",
    `1` = "Abandono"))
```



Distribución del abandono y la permanencia estudiantil según sexo.

Si miramos la distribución por sexo encontramos que no hay diferencias importantes entre los estudiantes que abandonan y los que no. En ambos casos, la composición de estudiantes es de aproximadamente el 50% de mujeres y 50% de hombres, por lo tanto, no parece existir evidencia de que el sexo sea un determinante del abandono estudiantil.